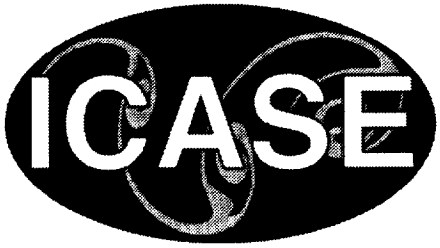
NASA/CR-1999-209099
ICASE Report No. 99-12



Approximation of the Newton Step by a Defect Correction Process

E. Arian
ICASE, Hampton, Virginia

A. Batterman and E.W. Sachs
Universität Trier, Trier Germany

Institute for Computer Applications in Science and Engineering
NASA Langley Research Center
Hampton, VA

Operated by Universities Space Research Association

National Aeronautics and
Space Administration

Langley Research Center
Hampton, Virginia 23681-2199

Prepared for Langley Research Center
under Contract NAS1-97046

February 1999

Available from the following:

NASA Center for AeroSpace Information (CASI)
7121 Standard Drive
Hanover, MD 21076-1320
(301) 621-0390

National Technical Information Service (NTIS)
5285 Port Royal Road
Springfield, VA 22161-2171
(703) 487-4650

APPROXIMATION OF THE NEWTON STEP BY A DEFECT CORRECTION PROCESS

E. ARIAN*, A. BATTERMANN†, AND E.W. SACHS‡

Abstract. In this paper, an optimal control problem governed by a partial differential equation is considered. The Newton step for this system can be computed by solving a coupled system of equations. To do this efficiently with an iterative defect correction process, a modifying operator is introduced into the system. This operator is motivated by local mode analysis. The operator can be used also for preconditioning in GMRES. We give a detailed convergence analysis for the defect correction process and show the derivation of the modifying operator. Numerical tests are done on the small disturbance shape optimization problem in two dimensions for the defect correction process and for GMRES.

Key words. optimal control governed by PDEs, iterative methods, defect correction, GMRES, preconditioning, Newton step, SQP.

Subject classification. Applied and Numerical Mathematics

1. Introduction. Many optimization problems can be formulated as equality constrained problems with a special structure. If one considers optimal control or optimal design problems, the variables are partitioned into the state and control or design variables which we denote by ϕ and u, respectively. This leads to the following problem formulation

$$\min_{(\phi,u)} \quad \mathcal{F}(\phi,u) \quad \text{s.t.} \quad h(\phi,u)=0.$$

If one is interested in algorithms with a fast rate of convergence, one would tend to use Newton's method for these problems. Note that this method can be applied in two different ways. Under appropriate assumptions, see Section 2.1, one can solve for each control variable u the system equation $h(\phi(u),u)=0$ to obtain a state $\phi(u)$ which depends on u. This is typical, when h represents a boundary value problem where the control variable is on the right hand side of the differential equation. Then one can apply Newton's method to the unconstrained minimization problem

$$\min_{u} \quad \mathcal{J}(u)=\mathcal{F}(\phi(u),u)\,,$$

where the step s is computed by solving the linear system

$$\mathcal{J}'(u)s=-\mathcal{J}(u)$$

for s.

*Institute for Computer Applications in Science and Engineering, Mail Stop 403, NASA Langley Research Center, Hampton, Virginia 23681 (arian@icase.edu). This research was supported by the National Aeronautics and Space Administration under NASA Contract No. NAS1-97046 while the author was in residence at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia 23681, and by the Stiftung Rheinland–Pfalz für Innovation and the Graduiertenkolleg Mathematische Optimierung of the Universität Trier while the author was in residence at the Universität Trier, 54286 Trier, Germany.

†Universität Trier, Fachbereich IV, Abteilung Mathematik, 54286 Trier, Germany (batt@uni-trier.de). This author was supported by the Stiftung Rheinland–Pfalz für Innovation and in part by the National Aeronautics and Space Administration under NASA Contract No. NAS1-97046 while the author was a visitor at the Institute for Computer Applications in Science and Engineering (ICASE), NASA Langley Research Center, Hampton, Virginia 23681.

‡Universität Trier, Fachbereich IV, Abteilung Mathematik, 54286 Trier, Germany (sachs@uni-trier.de).

Alternatively, one can keep all the variables (ϕ, u) and consider the necessary optimality conditions for the constrained optimization problem. Then one obtains the nonlinear equation in (ϕ, u, λ),

$$\mathcal{G}(\phi, u, \lambda) = \begin{pmatrix} h(\phi, u) \\ \mathcal{F}_\phi(\phi, u) + h_\phi^\times(\phi, u)\lambda \\ \mathcal{F}_u(\phi, u) + h_u^\times(\phi, u)\lambda \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix}.$$

This equation can be solved by Newton's method. For the step one has to solve the linear system

$$\mathcal{G}'(\phi, u, \lambda)(v, s, w)^T = -\mathcal{G}(\phi, u, \lambda).$$

This approach is the same if one applies an sequential quadratic programming method (SQP) to the constrained problem using the Newton multiplier update for the Lagrange multiplier, see [28].

It is important to note, that in both approaches at each step of Newton's method a linear system of equations has to be solved which exhibits the same structure for both cases, see Section 2.3. Only the right hand sides differ in these cases. If we denote the variables for the linear system by (v, s, w) then one obtains with the Lagrangian

$$\mathcal{L}(\phi, u, \lambda) = \mathcal{F}(\phi, u) + \lambda^\times\ h(\phi, u),$$

the linear system

$$\begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & h_\phi^\times \\ \mathcal{L}_{u\phi} & \mathcal{L}_{uu} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix} \begin{pmatrix} v \\ s \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ -\mathcal{L}_u \\ 0 \end{pmatrix} \quad \text{or} \quad - \begin{pmatrix} \mathcal{L}_\phi \\ \mathcal{L}_u \\ h \end{pmatrix}.$$

Since the solver of these linear systems often requires the largest part of the CPU time of an algorithm, it is the goal to utilize the special structure of the linear system in the linear system solver. This has been considered by [5] where several preconditioners were used and compared numerically and theoretically. Further discussions can be found in [10], [22], [26], [27], [18], [17], and [9]. In [16] the authors use a multilevel technique on the necessary optimality conditions in connection with Newton's method under box constraints on the control. The structure of the Newton system for optimal control problems is exploited in [15] to design special quasi Newton updates for problems including differential equations. It is well known that one does not need to solve the Newton equations exactly at each step [8] but only needs to decrease the accuracy in the residual as one approaches a stationary point of the optimization problem. This is another reason why we investigate iterative solvers for the Newton equation.

The gradient of $\mathcal{J}$ can be computed sequentially, see Section 2.9, by solving an adjoint equation after solving the nonlinear state equation. The system for the Newton step cannot be solved sequentially, since its variables (v, s, w) are coupled through the equations. This makes a Newton step quite expensive because an iterative procedure has to be applied. In some applications the variables ϕ and u are separated in $\mathcal{F}$ and h in such a way that the mixed terms in the second derivative of the Lagrangian disappear, i.e.,

$$\mathcal{L}_{\phi u} = \mathcal{L}_{u\phi} = 0.$$

If one would omit the upper left term $\mathcal{L}_{\phi\phi}$, then the linear system matrix has the form

$$\begin{pmatrix} 0 & 0 & h_\phi^\times \\ 0 & \mathcal{L}_{uu} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix}.$$

The resulting approximate Newton equation has the advantage that it can be solved sequentially, see Lemma 3.5. Hence it can be applied in an iterative way to improve the accuracy of the solution of the Newton step. If one analyzes its convergence, one obtains, see Section 3.3.4, that convergence is obtained if

$$\rho(\mathcal{L}_{uu}^{-1}\mathcal{H} - I) < 1\,,$$

where $\mathcal{H}$ denotes the Hessian of $\mathcal{J}$ or the reduced Hessian of the constrained optimization problem in terms of $\mathcal{F}$ and h. Thus, $\mathcal{H}$ is given by

$$\mathcal{H} = h_u^{\times}\, h_\phi^{-\times}\, \mathcal{L}_{\phi\phi}\, h_\phi^{-1}\, h_u - h_u^{\times}\, h_\phi^{-\times}\, \mathcal{L}_{\phi u} - \mathcal{L}_{u\phi}\, h_\phi^{-1}\, h_u + \mathcal{L}_{uu}\,. \tag{1.1}$$

Since the condition $\rho(\mathcal{L}_{uu}^{-1}\mathcal{H} - I) < 1$ might be too restrictive we investigate the following strategy in this paper.

At first we replace $\mathcal{L}_{uu}$ by a term $\mathcal{L}_{u,\epsilon} = \mathcal{L}_{uu}(I + \epsilon\mathcal{P})$, where $\mathcal{P}$ and ϵ can be chosen properly. This choice in general depends on the application under consideration. Then the system equation changes and we consider a separation into an outer loop iteration and an inner loop iteration. The outer loop iteration is given by

$$\begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & h_\phi^{\times} \\ \mathcal{L}_{u\phi} & \mathcal{L}_{u,\epsilon} & h_u^{\times} \\ h_\phi & h_u & 0 \end{pmatrix} \begin{pmatrix} v \\ s \\ w \end{pmatrix}^{n+1} = \begin{pmatrix} 0 \\ -\mathcal{L}_u + \mathcal{L}_{uu}\epsilon\mathcal{P}s^n \\ 0 \end{pmatrix}, \tag{1.2}$$

which is solved by an inner loop iteration through

$$\begin{pmatrix} 0 & 0 & h_\phi^{\times} \\ 0 & \mathcal{L}_{u,\epsilon} & h_u^{\times} \\ h_\phi & h_u & 0 \end{pmatrix} x^{k+1} = r - \begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & 0 \\ \mathcal{L}_{u\phi} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix} x^k\,. \tag{1.3}$$

Here, r denotes the right hand side of the original linear system equation. Thus the advantage of the sequential solution of the approximate Newton step is retained. In Section 3.3 we analyze the convergence properties of this iterative solver for the Newton step.

The choice of the operator $\mathcal{P}$ in the iteration (1.2), (1.3) is crucial for the convergence properties of the resulting scheme. We suggest to make this choice by analyzing the optimization problem on the infinite dimensional level, i.e., before discretization is applied. The operator $\mathcal{P}$ can be derived approximately with local mode (Fourier) analysis of the reduced Hessian, following [2], [3]. By that we are using the structure induced by the governing partial differential equation (PDE) to accelerate the convergence process. Once $\mathcal{P}$ has been determined, it can be applied in various manners to accelerate the convergence process, for example by taking the iteration matrix in (1.3) as a preconditioner for Krylov subspace methods. However, we concentrate on the defect correction process (1.2), (1.3) because it is simple to apply. Once a code is given for the solution of the state and costate equations, the implementation of the defect correction process is reduced to successively solving the linearized state and the costate equations, with possibly different right hand sides. The defect correction process is formally introduced in Section 3.

In Section 4.1 we apply this approach to an optimal shape design problem arising in aerodynamics. It is a boundary control problem in which the shape of the solid wall is optimized by modifying the right hand side of the Neumann boundary condition. It involves the solution of an elliptic boundary value problem in two space dimensions. By using a local mode analysis of the reduced Hessian (1.1) which is done in Section 4.7, we obtain an operator $\mathcal{P}$ and an indication of the convergence properties of the method for a small step size of the discretization. This is verified by the numerical results and the rates obtained for the example. In Section 4.10 we present numerical results for the defect correction process and for preconditioned GMRES using different choices of the operator $\mathcal{P}$. We obtain a significant decrease of the residual in the first iterations. This convergence property is crucial in many applications that involve computationally intensive cost functional evaluations and derivative computations.

2. General Approach. In this section, a general equality constrained optimal control problem is addressed. The necessary optimality conditions are given for this problem together with the optimality conditions for an equivalent unconstrained problem.

2.1. Problem Formulation. We repeat the problem to be considered in its general formulation

$$\min_{(\phi,u)} \quad \mathcal{F}(\phi,u) \quad \text{s.t.} \quad h(\phi,u)=0\,. \tag{2.1}$$

The constraint $h(\phi,u)=0$ denotes the state equation where ϕ is the state variable and u is the control, or design, variable. Under the following assumption, the equation can be solved uniquely in ϕ for a given u. Also, the Newton step for the minimization problem (2.1) is well defined. In our presentation we follow [19].

ASSUMPTION 2.1. *Let $\mathcal{X},\mathcal{Y},\mathcal{Z}$ be Hilbert spaces. Let $\mathcal{F}$: $\mathcal{X}\times\mathcal{Y}\to\mathbb{R}$ and h: $\mathcal{X}\times\mathcal{Y}\to\mathcal{Z}$ be twice continuously Fréchet differentiable. Let h_ϕ, the partial Fréchet derivative of h with respect to ϕ, be bijective and continuous. Let h_u, the partial Fréchet derivative of h with respect to u, be continuous.*

REMARK 2.2. *By Assumption* 2.1, *the inverse of $h_\phi(\phi_c,u_c)$ at the point (ϕ_c,u_c) exists. The derivative $h_\phi(\phi_c,u_c)(\delta\phi)$ of h with respect to ϕ at the point (ϕ_c,u_c) is linear in the increment $\delta\phi$. Thus, the inverse of $h_\phi^\times(\phi_c,u_c)$ also exists and is continuous (see, e.g.,* [7]*). Here and in the following, the superscript $^\times$ denotes the adjoint operator or space. We will in the following often denote $h_\phi(\phi_c,u_c)$ by h_ϕ and apply similar conventions to other functions. Note also that the relationship $(h_\phi^\times)^\times = h_\phi$ holds for h_ϕ and the other considered functions (see, e.g.,* [7]*).*

The implicit function theorem (see, e.g., [30, p.150]) allows to define the following mapping ϕ.

LEMMA 2.3. *Let Assumption* 2.1 *hold and let ϕ_c,u_c satisfy $h(\phi_c,u_c)=0$. Let $\mathcal{U}$ be an open neighborhood of $(\phi_c,u_c)\in\mathcal{X}\times\mathcal{Y}$. Then there exists a unique mapping ϕ: $\mathcal{U}\to\mathcal{X}$ that is twice continuously Fréchet differentiable in a neighborhood $\mathcal{U}$ of $u_c\in\mathcal{Y}$ and satisfies $h(\phi(u),u)=0 \quad \forall u\in\mathcal{U}$. Furthermore, the derivative ϕ' of ϕ with respect to u is given by*

$$\phi'(u) = -h_\phi^{-1}(\phi(u),u)\; h_u(\phi(u),u) \quad \forall u\in\mathcal{U}. \tag{2.2}$$

In the situation of Lemma 2.3 we can define the following unconstrained optimization problem which is equivalent to the problem in its original formulation (2.1):

$$\min_u \quad \mathcal{J}(u)=\mathcal{F}(\phi(u),u). \tag{2.3}$$

2.2. The Necessary Optimality Conditions. The Lagrangian function for problem (2.1) is given by

$$\mathcal{L}(\phi,u,\lambda)=\mathcal{F}(\phi,u)+\lambda^\times\; h(\phi,u), \tag{2.4}$$

where λ denotes the Lagrange multiplier defined in $\mathcal{Z}^\times$, the dual of $\mathcal{Z}$. The first order necessary optimality conditions for a minimizer of problem (2.1) are given (see, e.g., [20]) by setting the gradient $\nabla\mathcal{L}$ of the Lagrangian function $\mathcal{L}$ to zero, i.e., by the equations

$$\text{state:} \quad \mathcal{L}_\lambda = h(\phi,u) \qquad = 0, \tag{2.5}$$

$$\text{costate:} \quad \mathcal{L}_\phi = \mathcal{F}_\phi + h_\phi^\times\,\lambda = 0, \tag{2.6}$$

$$\text{design:} \quad \mathcal{L}_u = \mathcal{F}_u + h_u^\times\,\lambda = 0. \tag{2.7}$$

The gradient, $g=\mathcal{J}'(u)$, is given by the following lemma (this is the necessary optimality condition of first order for the unconstrained problem (2.3) (see, e.g., [20])).

LEMMA 2.4. *Let Assumption* 2.1 *hold. Define for* $u \in \mathcal{Y}$

i.) *a function* $\phi(u)$ *that satisfies* (2.5) *and*

ii.) *a function* $\lambda(u)$ *as the unique solution of the adjoint equation* (2.6)*, i.e., of*

$$h_\phi^\times(\phi(u),u)\ \lambda(u) = -\mathcal{F}_\phi(\phi(u),u). \tag{2.8}$$

Then the gradient g *of* (2.3) *is given by*

$$g = \mathcal{J}'(u) = \mathcal{F}_u(\phi(u),u) + h_u^\times(\phi(u),u)\ \lambda(u). \tag{2.9}$$

Proof. The assertion follows from the chain rule and (2.2):

$$\begin{aligned} g \equiv \mathcal{J}'(u) &= \mathcal{F}_u(\phi(u),u) + \phi'(u)^\times\, \mathcal{F}_\phi(\phi(u),u) \\ &= \mathcal{F}_u(\phi(u),u) + h_u^\times(\phi(u),u)\,(-h_\phi^{-\times}(\phi(u),u))\,\mathcal{F}_\phi(\phi(u),u). \end{aligned}$$

□

2.3. The Newton Step. The definition of the Newton step for the unconstrained problem (2.3) requires the computation of the second derivative of the objective functional $\mathcal{J}$. This, in turn, requires the differentiation of the adjoint variable λ.

LEMMA 2.5. *Let Assumption* 2.1 *hold. Then* $\lambda(u)$ *defined in the adjoint equation* (2.8) *is differentiable, and the derivative is given by*

$$\lambda'(u) = -h_\phi^{-\times}(\phi(u),u)\ [\mathcal{L}_{\phi u}(\phi(u),u,\lambda(u)) + \phi'(u)\ \mathcal{L}_{\phi\phi}(\phi(u),u,\lambda(u))]\,. \tag{2.10}$$

Proof. Define a map κ by $\kappa(\lambda,u) = \mathcal{L}_\phi(\phi(u),u,\lambda)$, where by (2.8) $\lambda = \lambda(u)$ solves $\kappa(\lambda(u),u) = 0$. Since $\kappa_\lambda(\lambda,u)(.) = h_\phi^\times(\phi(u),u)(.)$ is invertible by Assumption 2.1 (see Remark 2.2), the assertion follows by the implicit function theorem. □

With this result we can now express the Hessian of the unconstrained problem (2.3).

THEOREM 2.6. *Let Assumption* 2.1 *hold. Then* $\mathcal{J}$ *defined in* (2.3) *is twice Fréchet differentiable, and the Hessian is given by*

$$\mathcal{H} = \mathcal{J}''(u) = h_u^\times\, h_\phi^{-\times}\, \mathcal{L}_{\phi\phi}\, h_\phi^{-1}\, h_u - h_u^\times\, h_\phi^{-\times}\, \mathcal{L}_{\phi u} - \mathcal{L}_{u\phi}\, h_\phi^{-1}\, h_u + \mathcal{L}_{uu}. \tag{2.11}$$

Proof. Apply the chain rule to equation (2.9) and use equations (2.2) and (2.10). □

With Hessian, $\mathcal{H}$, and gradient, g, for problem (2.3) the Newton step, s, is given by

$$\mathcal{H}\, s = -g. \tag{2.12}$$

REMARK 2.7. *So far we assumed that the state and costate equations are feasible at every SQP step (reduced SQP). Thus the Newton equation is given by* (2.12)*. In case of a full SQP algorithm, where feasibility is not required at each step,* (2.12) *has to be modified to*

$$\mathcal{H}\, s = -\mathcal{L}_u - \mathcal{L}_{u\phi}\, h_\phi^{-1} h + h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi\phi}\, h_\phi^{-1} h - h_u^\times h_\phi^{-\times} \mathcal{L}_\phi. \tag{2.13}$$

The right hand side of (2.13) *consists of the residual of the design equation (this term is equal to the reduced gradient,* g*, when the state and costate equations are solved), two terms that vanish when feasibility is achieved, i.e., when* $h(\phi,u) = 0$*, and the last term that vanishes when the costate equation is feasible, i.e., when* $\mathcal{L}_\phi(\phi,u,\lambda) = 0$.

THEOREM 2.8. *Let Assumption* 2.1 *hold. The Newton step, s, defined by* (2.12) *can be computed by solving the self–adjoint system of equations*

$$\begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & h_\phi^\times \\ \mathcal{L}_{u\phi} & \mathcal{L}_{uu} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix} \begin{pmatrix} v \\ s \\ w \end{pmatrix} = \begin{pmatrix} 0 \\ -g \\ 0 \end{pmatrix}. \tag{2.14}$$

REMARK 2.9. *In case of a full SQP on a nonlinear problem, the right hand side of* (2.14) *should be modified to $-(\mathcal{L}_\phi, \mathcal{L}_u, \mathcal{L}_\lambda)^T$ to be consistent with* (2.13).

Proof. For $s \in \mathcal{Y}$ define $v \in \mathcal{X}$ and $w \in \mathcal{Z}^\times$ by

$$\begin{aligned} v &= -h_\phi^{-1} h_u s, \\ w &= -h_\phi^{-\times} (\mathcal{L}_{\phi\phi} v + \mathcal{L}_{\phi u} s). \end{aligned} \tag{2.15}$$

Then by (2.11) and (2.12) the Newton step, s, satisfies the equation

$$\begin{aligned} -g &= h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi\phi} h_\phi^{-1} h_u s - h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi u} s - \mathcal{L}_{u\phi} h_\phi^{-1} h_u s + \mathcal{L}_{uu} s \\ &= \mathcal{L}_{u\phi} v + \mathcal{L}_{uu} s + h_u^\times (-h_\phi^{-\times} \mathcal{L}_{\phi\phi} v - h_\phi^{-\times} \mathcal{L}_{\phi u} s). \end{aligned}$$

□

We have shown that solving (2.12) for the Newton step, s, is equivalent to solving (2.14). Writing the right hand side of (2.14), $(0, -g, 0)^T$, as $(r_1, r_2, r_3)^T = r$, we denote (2.14) by

$$\mathcal{K} x = r, \tag{2.16}$$

and the exact solution of (2.16) by x^*. Thus, $\mathcal{K}$ is defined as

$$\mathcal{K} = \begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & h_\phi^\times \\ \mathcal{L}_{u\phi} & \mathcal{L}_{uu} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix}. \tag{2.17}$$

NOTATION 2.10. *In the following, the vector of unknowns $(v, s, w)^T$ will often be referred to as x. For the description of iterative processes, the superscripts $^{c,+}$ will denote current and new iterates, respectively, e.g., x^c the current x–iterate. The solution of an iterative process will be indicated with the superscript *, e.g., x^*. In addition, the error in the vector x is denoted by e; specifically, e^c is the error in the current x–iterate. The error in the components, e.g. s, will be denoted by e_s. Thus, for instance,*

$$x^c - x^* = \begin{pmatrix} v^c \\ s^c \\ w^c \end{pmatrix} - \begin{pmatrix} v^* \\ s^* \\ w^* \end{pmatrix} = \begin{pmatrix} e_v^c \\ e_s^c \\ e_w^c \end{pmatrix} = e^c.$$

3. The Solution Method. In order to solve for the Newton step, a defect correction process is employed. The defect correction process (see [12], [24]) is derived in this section. Convergence of the process is governed by the choice of approximating operator $\tilde{\mathcal{K}}$. A detailed discussion and convergence analysis is done in this section.

3.1. The Defect Correction Process. Solving for the Newton step, i.e., solving (2.12), is equivalent to solving $\mathcal{K} x = r$ in (2.16). The idea of the defect correction approach is to replace $\mathcal{K}$ by a simple approximation $\tilde{\mathcal{K}}$. The solution of the approximate problem

$$\tilde{\mathcal{K}} x = \tilde{r} \tag{3.1}$$

is then reached iteratively. It is essential that $\tilde{\mathcal{K}}^{-1}$ be relatively simple, i.e., that it is much easier to find a solution to (3.1) than to (2.16).

We now introduce the defect correction process.

ASSUMPTION 3.1. *Guided by the treatment in* [12], *we assume the following.*

i.) *Let* $\mathcal{K} : \mathcal{E} \supset \mathcal{D} \to \hat{\mathcal{D}} \subset \hat{\mathcal{E}}$ *continuous and bijective,* $\mathcal{E}$, $\hat{\mathcal{E}}$ *Hilbert spaces,* $\mathcal{D}$, $\hat{\mathcal{D}}$ *closed subsets.*

ii.) *The defect*

$$d(\tilde{x}) = \mathcal{K}\tilde{x} - \tilde{r} \tag{3.2}$$

can be evaluated for approximate solutions $\tilde{x} \in \mathcal{D}$ *to all neighboring problems. The neighboring problem is to find* $\tilde{x} \in \mathcal{D}$ *with* $\mathcal{K}\tilde{x} = \tilde{r}$ *for given* $\tilde{r} \in \hat{\mathcal{D}}$.

iii.) *The approximate problem* (3.1) *can be solved uniquely for* $\tilde{r} \in \hat{\mathcal{D}}$, *i.e., we assume the existence of an approximative inverse* $\tilde{\mathcal{K}}^{-1}$ *of* $\mathcal{K}$ *such that* $\tilde{\mathcal{K}}^{-1}\mathcal{K}x \approx x$ *for* $x \in \mathcal{D}$ *and* $\mathcal{K}\tilde{\mathcal{K}}^{-1}\tilde{r} \approx \tilde{r}$ *for* $\tilde{r} \in \hat{\mathcal{D}}$.

Assuming that we know an approximation $x^c \in \mathcal{D}$ for x^* and that we have computed its defect $d(x^c) = \mathcal{K}x^c - r = \mathcal{K}x^c - \mathcal{K}x^*$, this information can be used for the computation of an update x^+ by means of solving a problem (3.1). The error $e^c = x^c - x^*$ satisfies $e^c = \mathcal{K}^{-1}(r + d(x^c)) - \mathcal{K}^{-1}r = \mathcal{K}^{-1}d(x^c)$. Instead of performing the difficult solve with $\mathcal{K}$, we use the approximation $\tilde{\mathcal{K}}$ to compute $\tilde{e}^c = \tilde{\mathcal{K}}^{-1}d(x^c)$ and use this quantity as a correction for x^c. The iterative usage leads to the scheme

$$x^+ = x^c + \tilde{e}^c = (I - \tilde{\mathcal{K}}^{-1}\mathcal{K})x^c + x^0 \tag{3.3}$$

with $x^0 = \tilde{\mathcal{K}}^{-1}r$ as initial iterate. We define $\mathcal{R}$ by the relation $\mathcal{R} := \mathcal{K} - \tilde{\mathcal{K}}$ and call $\tilde{\mathcal{K}} + \mathcal{R}$ a *splitting* of $\mathcal{K}$. With this notation we write (3.3) as

$$\tilde{\mathcal{K}}x^+ = r - \mathcal{R}x^c \tag{3.4}$$

to indicate that we do not really apply $\tilde{\mathcal{K}}^{-1}$ but solve the system with $\tilde{\mathcal{K}}$.

3.2. The Modified System Defect Correction. We will in the following apply the defect correction process not precisely in the way it was derived in the above Section 3.1, but introduce two changes. First, the defect correction process described in Section 3.1 can be nested, i.e., an inner defect correction loop can be used to find the solution to the system (3.4) that has to be solved in each step of an outer loop. This is the point of view we take in the following presentation. Secondly, we modify the system (2.14) that we are interested in solving with the help of an additional operator $\mathcal{P}$. We call this approach the modified system defect correction process.

We now turn to the splittings we propose for the solution of (2.14) by the process (3.4). The choice of $\tilde{\mathcal{K}}$ in the splitting $\mathcal{K} = \tilde{\mathcal{K}} + \mathcal{R}$ is crucial for both applicability and convergence of the iterative scheme (3.4). Our choice of $\tilde{\mathcal{K}}$ is motivated by the structure of the underlying system (2.14). We now supplement Assumption 2.1.

ASSUMPTION 3.2. *Let* $\mathcal{L}_{uu} \neq 0, \mathcal{L}_{uu}^{-1}$ *exist.*

We will in the following see that the part $\mathcal{L}_{uu}$ in the system $\mathcal{K}$ is crucial for the performance of the solution method. The convergence requirement is

$$\rho(\mathcal{L}_{uu}^{-1}\mathcal{H} - I) < 1.$$

In general, $\mathcal{L}_{uu}$ will not model the Hessian, $\mathcal{H}$, very well, and convergence is not necessarily ascertained. To ensure convergence, and to allow for convergence acceleration, we introduce an operator $\mathcal{P}$, which will be used to modify the system. The choice of $\mathcal{P}$ will be discussed subsequent to the convergence analysis in Section 3.3.

ASSUMPTION 3.3. *Let an operator* $\mathcal{P}$ *on* $\mathcal{Y}$ *and a real scalar* ϵ *exist such that* $I + \epsilon\mathcal{P}$ *is invertible.*

Here, I is the identity on $\mathcal{Y}$. Note that Assumption 3.3 is necessarily satisfied for small ϵ. The operator $\mathcal{P}$ will in general be applied to modify $\mathcal{L}_{uu}$. As an abbreviation we use

$$\mathcal{L}_{u,\epsilon} = \mathcal{L}_{uu}\,(I + \epsilon\mathcal{P}).$$

Thus, $\mathcal{L}_{uu}$ is naturally regained as $\mathcal{L}_{u,0}$.

3.2.1. Outer Loop. To solve the system $\mathcal{K}\,x = r$ in (2.16) efficiently, we modify $\mathcal{K}$, $\mathcal{K}$ given in (2.17), through replacing $\mathcal{L}_{uu}$ by $\mathcal{L}_{u,\epsilon} = \mathcal{L}_{uu}\,(I + \epsilon\mathcal{P})$ to

$$\tilde{\mathcal{K}}_O = \begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & h_\phi^\times \\ \mathcal{L}_{u\phi} & \mathcal{L}_{u,\epsilon} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix}. \tag{3.5}$$

With

$$\mathcal{R}_O = \begin{pmatrix} 0 & 0 & 0 \\ 0 & -\mathcal{L}_{uu}\epsilon\mathcal{P} & 0 \\ 0 & 0 & 0 \end{pmatrix}, \tag{3.6}$$

the equality $\tilde{\mathcal{K}}_O + \mathcal{R}_O = \mathcal{K}$ holds so that $\tilde{\mathcal{K}}_O$, $\mathcal{R}_O$ define a splitting of $\mathcal{K}$.

For the solution of $\mathcal{K}\,x = r$ we start a defect correction process given by

$$\tilde{\mathcal{K}}_O\,x_O^{n+1} = r - \mathcal{R}_O\,x_O^n\,, \tag{3.7}$$

which is equivalent to

$$\begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & h_\phi^\times \\ \mathcal{L}_{u\phi} & \mathcal{L}_{u,\epsilon} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix} \begin{pmatrix} v \\ s \\ w \end{pmatrix}^{n+1} = \begin{pmatrix} 0 \\ -g_0 + \mathcal{L}_{uu}\epsilon\mathcal{P}s^n \\ 0 \end{pmatrix}. \tag{3.8}$$

We start at $n = 0$ with a starting point v^0, s^0, w^0. In each step of this defect correction process, a linear system has to be solved. One possibility is to do this via an inner defect correction loop.

3.2.2. Inner Loop. For the inner loop we define a splitting of the system matrix $\tilde{\mathcal{K}}_O$ in (3.8). The splitting $\tilde{\mathcal{K}}_O = \tilde{\mathcal{K}}_I + \mathcal{R}_I$ we propose is given by

$$\tilde{\mathcal{K}}_I = \begin{pmatrix} 0 & 0 & h_\phi^\times \\ 0 & \mathcal{L}_{u,\epsilon} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix}, \quad \mathcal{R}_I = \begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & 0 \\ \mathcal{L}_{u\phi} & 0 & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.9}$$

The corresponding inner loop is

$$\tilde{\mathcal{K}}_I\,x_I^{k+1} = r_O - \mathcal{R}_I\,x_I^k\,, \tag{3.10}$$

where r_O is the right hand side of the outer loop, i.e., r_O is given by $r_O = r - \mathcal{R}_O\,x_O^n$. This amounts in each step to

$$\begin{pmatrix} 0 & 0 & h_\phi^\times \\ 0 & \mathcal{L}_{u,\epsilon} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}^{k+1} = \begin{pmatrix} -\mathcal{L}_{\phi\phi}\,x_1^k - \mathcal{L}_{\phi u}\,x_2^k \\ -g_0 + \mathcal{L}_{uu}\epsilon\mathcal{P}s^n - \mathcal{L}_{u\phi}x_1^k \\ 0 \end{pmatrix}. \tag{3.11}$$

Starting at $k = 0$, use $x_I^0 = x_O^n$. Set the solution x_I^* of the inner loop as the new outer loop iterate x_O^{n+1}.

REMARK 3.4. *For the splitting defined by* (3.9), *the iterative scheme* (3.4) *can be viewed as applying (forward) Gauss–Seidel on the system*

$$\begin{pmatrix} h_\phi^\times & \mathcal{L}_{\phi u} & \mathcal{L}_{\phi\phi} \\ h_u^\times & \mathcal{L}_{u,\epsilon} & \mathcal{L}_{u\phi} \\ 0 & h_u & h_\phi \end{pmatrix}. \tag{3.12}$$

Because (3.12) *is, even for* $\epsilon = 0$, *a non–selfadjoint permutation of system* (2.17), *symmetric Gauss–Seidel (i.e., forward Gauss–Seidel followed by backward Gauss–Seidel, see* [11], [14]*), does not lead to a selfadjoint operator* $\mathcal{M} = I - \tilde{\mathcal{K}}^{-1}\mathcal{K}$ *in* (3.3).

3.2.3. One Inner Iteration. If only one inner iteration is performed, inner and outer loop can be combined in one closed formula

$$\tilde{\mathcal{K}}\, x^{k+1} = r - \mathcal{R}\, x^k. \tag{3.13}$$

This is described by (3.11) with s^n replaced by x_2^k. The corresponding splitting is $\mathcal{K} = \tilde{\mathcal{K}} + \mathcal{R}$ with $\tilde{\mathcal{K}} = \tilde{\mathcal{K}}_I$, $\mathcal{R} = \mathcal{R}_I + \mathcal{R}_O$, and given by

$$\tilde{\mathcal{K}} = \begin{pmatrix} 0 & 0 & h_\phi^\times \\ 0 & \mathcal{L}_{u,\epsilon} & h_u^\times \\ h_\phi & h_u & 0 \end{pmatrix}, \quad \mathcal{R} = \begin{pmatrix} \mathcal{L}_{\phi\phi} & \mathcal{L}_{\phi u} & 0 \\ \mathcal{L}_{u\phi} & -\mathcal{L}_{uu}\,\epsilon\mathcal{P} & 0 \\ 0 & 0 & 0 \end{pmatrix}. \tag{3.14}$$

3.2.4. Applicability. With the splittings defined in the preceding Sections 3.2.2 and 3.2.3, the solution x^+ of the scheme (3.4), i.e., to $\tilde{\mathcal{K}}\, x^+ = r - \mathcal{R}\, x^c$, can be computed at the cost of solving three linear subsystems.

LEMMA 3.5. *Let the Assumptions* 2.1, 3.2 *and* 3.3 *hold. The solution to the systems* (3.10) *and* (3.13) *can be accomplished at the cost of solving three linear subsystems.*

Proof. Since the systems are blocktriangular, back–substitution furnishes the solution.

1.) The solution x^{k+1} to (3.13), i.e., to $\tilde{\mathcal{K}}\, x^{k+1} = r - \mathcal{R}\, x^k$, can be computed by successively solving the systems

$$\begin{aligned} h_\phi^\times\, x_3^{k+1} &= r_1 - \mathcal{L}_{\phi\phi}\, x_1^k - \mathcal{L}_{\phi u}\, x_2^k, \\ \mathcal{L}_{u,\epsilon}\, x_2^{k+1} &= r_2 - \mathcal{L}_{u\phi}\, x_1^k + \mathcal{L}_{uu}\,\epsilon\mathcal{P} x_2^k - h_u^\times\, x_3^{k+1}, \\ h_\phi\, x_1^{k+1} &= r_3 - h_u\, x_2^{k+1}. \end{aligned}$$

2.) The solution x_I^{k+1} to (3.10), i.e., to $\tilde{\mathcal{K}}_I\, x_I^{k+1} = r_O - \mathcal{R}_I\, x_I^k$, can be computed by successively solving the systems

$$\begin{aligned} h_\phi^\times\, x_{I,3}^{k+1} &= r_{O,1} - \mathcal{L}_{\phi\phi}\, x_{I,1}^k - \mathcal{L}_{\phi u}\, x_{I,2}^k, \\ \mathcal{L}_{u,\epsilon}\, x_{I,2}^{k+1} &= r_{O,2} - \mathcal{L}_{u\phi}\, x_{I,1}^k - h_u^\times\, x_{I,3}^{k+1}, \\ h_\phi\, x_{I,1}^{k+1} &= r_{O,3} - h_u\, x_{I,2}^{k+1}. \end{aligned}$$

□ We now turn to the convergence analysis of the iteration (3.4) for the proposed splittings.

3.3. Convergence Analysis. The convergence of the defect correction process depends on properties of the iteration operator $\mathcal{M} = I - \tilde{\mathcal{K}}^{-1}\mathcal{K} = -\tilde{\mathcal{K}}^{-1}\mathcal{R}$ in the process (3.3). First, we state the basic requirement on $\mathcal{M}$ in Theorem 3.6. We then address the necessary and sufficient conditions for convergence of the modified system defect correction for the processes described in Sections 3.2.1, 3.2.2, and 3.2.3. For this, we investigate the convergence–governing matrices $\mathcal{M}_O$, $\mathcal{M}_I$, and $\mathcal{M}$ of the processes. The composing operators are derived from the respective splittings defined in (3.5) and (3.6), (3.9) and (3.14).

The basic convergence requirement on $\mathcal{M}$ is described in [30, Cor.1.13].

THEOREM 3.6. *Let Assumptions* 2.1, 3.1, *and* 3.2 *hold. If the condition*

$$\rho(\mathcal{M}) < 1 \tag{3.15}$$

holds for the spectral radius ρ of $\mathcal{M} = I - \tilde{\mathcal{K}}^{-1}\mathcal{K}$ in process (3.3), *then the iteration converges, for every r and for arbitrary initial element x^0, to a unique solution x^* of* (2.16).

In the following we often need the invertibility of the reduced Hessian, $\mathcal{H}$, and of a modification, $\mathcal{H}_\epsilon = \mathcal{H} + \mathcal{L}_{uu}\epsilon\mathcal{P}$. The invertibility is guaranteed for small ϵ under the usual second order sufficiency conditions for optimization problems because these require, with some constant c,

$$\mathcal{L}_{((\phi,u),(\phi,u))}(\phi^*,u^*,\lambda^*)(\delta\phi,\delta u)(\delta\phi,\delta u) \geq c\,\|(\delta\phi,\delta u)\|^2 \tag{3.16}$$

for all those $(\delta\phi,\delta u)$ that satisfy $h_\phi\,\delta\phi + h_u\,\delta u = 0$. The last condition allows to substitute $\delta\phi = -h_\phi^{-1}\,h_u\,\delta u$, thus leading to

$$\begin{aligned} &h_u^\times\, h_\phi^{-\times}\mathcal{L}_{\phi\phi}\,h_\phi^{-1}h_u(\delta u,\delta u) - h_u^\times\, h_\phi^{-\times}\mathcal{L}_{\phi u}(\delta u,\delta u) - \mathcal{L}_{u\phi}h_\phi^{-1}h_u(\delta u,\delta u) + \mathcal{L}_{uu}(\delta u,\delta u)\\ &= \mathcal{H}(\delta u,\delta u) \geq c\,(\|h_\phi^{-1}\,h_u\,\delta u\|^2 + \|\delta u\|^2) \geq c\,\|\delta u\|^2. \end{aligned}$$

Therefore, $\mathcal{H}$ is invertible and so is $\mathcal{H}_\epsilon$ for small ϵ.

ASSUMPTION 3.7. *Let a real scalar $\bar{\epsilon} > 0$ exist such that $\mathcal{H}_\epsilon = \mathcal{H} + \mathcal{L}_{uu}\epsilon\mathcal{P}$ is invertible for $0 \leq \epsilon < \bar{\epsilon}$.*

In addition we will need the following lemma to prove the central statements in Theorems 3.9, 3.10, and 3.11.

LEMMA 3.8. *Let Assumptions* 2.1 *and* 3.7 *hold. Then the inverse of the operator $\mathcal{K}$ defined in* (2.17), *written as*

$$\mathcal{K}^{-1} = \begin{pmatrix} \mathcal{K}_{11}^{-1} & \mathcal{K}_{12}^{-1} & \mathcal{K}_{13}^{-1} \\ \mathcal{K}_{12}^{-\times} & \mathcal{K}_{22}^{-1} & \mathcal{K}_{23}^{-1} \\ \mathcal{K}_{13}^{-\times} & \mathcal{K}_{23}^{-\times} & \mathcal{K}_{33}^{-1} \end{pmatrix},$$

is given by its entries

$$\begin{aligned} \mathcal{K}_{11}^{-1} &= h_\phi^{-1}h_u\mathcal{H}^{-1}h_u^\times h_\phi^{-\times}, \\ \mathcal{K}_{12}^{-1} &= h_\phi^{-1}h_u\mathcal{H}^{-1}, \\ \mathcal{K}_{13}^{-1} &= h_\phi^{-1} - h_\phi^{-1}h_u\mathcal{H}^{-1}h_u^\times h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1} + h_\phi^{-1}h_u\mathcal{H}^{-1}\mathcal{L}_{u\phi}h_\phi^{-1}, \\ \mathcal{K}_{22}^{-1} &= \mathcal{H}^{-1}, \\ \mathcal{K}_{23}^{-1} &= \mathcal{H}^{-1}h_u^\times h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1} - \mathcal{H}^{-1}\mathcal{L}_{u\phi}h_\phi^{-1}, \\ \mathcal{K}_{33}^{-1} &= -h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1} + h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1}h_u\mathcal{H}^{-1}h_u^\times h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1} + h_\phi^{-\times}\mathcal{L}_{\phi u}\mathcal{H}^{-1}\mathcal{L}_{u\phi}h_\phi^{-1} \\ &\quad -h_\phi^{-\times}\mathcal{L}_{\phi u}\mathcal{H}^{-1}h_u^\times h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1} - h_\phi^{-\times}\mathcal{L}_{\phi\phi}h_\phi^{-1}h_u\mathcal{H}^{-1}\mathcal{L}_{u\phi}h_\phi^{-1}. \end{aligned}$$

Here, $\mathcal{H}$, explicitly given in (2.11), *denotes the reduced Hessian of the constrained problem* (2.1). *For the operator $\tilde{\mathcal{K}}_O$ in* (3.5) *which differs from $\mathcal{K}$ only in the central entry $\mathcal{L}_{u,\epsilon}$, the inverse is given by*

$$\tilde{\mathcal{K}}_O^{-1} = \begin{pmatrix} \mathcal{K}_{11,\epsilon}^{-1} & \mathcal{K}_{12,\epsilon}^{-1} & \mathcal{K}_{13,\epsilon}^{-1} \\ \mathcal{K}_{12,\epsilon}^{-\times} & \mathcal{K}_{22,\epsilon}^{-1} & \mathcal{K}_{23,\epsilon}^{-1} \\ \mathcal{K}_{13,\epsilon}^{-\times} & \mathcal{K}_{23,\epsilon}^{-\times} & \mathcal{K}_{33,\epsilon}^{-1} \end{pmatrix},$$

where the entries differ from those of $\mathcal{K}^{-1}$ insofar as $\mathcal{H}$ is replaced by $\mathcal{H}_\epsilon = \mathcal{H} + \mathcal{L}_{uu}\epsilon\mathcal{P} = h_u^\times\, h_\phi^{-\times}\,\mathcal{L}_{\phi\phi}\,h_\phi^{-1}\,h_u - h_u^\times\, h_\phi^{-\times}\,\mathcal{L}_{\phi u} - \mathcal{L}_{u\phi}\,h_\phi^{-1}\,h_u + \mathcal{L}_{uu}(I + \epsilon\mathcal{P})$.

3.3.1. Convergence Analysis of the Outer Loop. We now investigate the convergence properties of the outer loop. The convergence of $x^n = (v^n, s^n, w^n)^T$ can be characterized by the convergence of the sequence s^n. This is important because in the application (see Section 4) x is a vector which consists of functions defined also on the whole two–dimensional domain whereas s is only defined on the boundary.

THEOREM 3.9. *Under Assumptions* 2.1, 3.2, 3.3 *and* 3.7 *the following statements hold.*

1.) *The iteration* (3.7) *converges if and only if the iteration*

$$s^{n+1} = \tilde{\mathcal{K}}_{22,\epsilon}^{-1} \mathcal{L}_{uu} \epsilon \mathcal{P} s^n + \tilde{r}_2 \tag{3.17}$$

converges, where $\tilde{r} = \tilde{\mathcal{K}}_O^{-1} r$.

2.) *If the spectral radius of* $I - (\mathcal{H} + \mathcal{L}_{uu}\, \epsilon \mathcal{P})^{-1}\, \mathcal{H}$, *defined on the boundary, satisfies*

$$\rho(I - (\mathcal{H} + \mathcal{L}_{uu}\, \epsilon \mathcal{P})^{-1}\, \mathcal{H}) < 1, \tag{3.18}$$

then the iterates of (3.7) *converge to the solution* x^* *of* (2.16).

Proof. Denoting $\tilde{\mathcal{K}}_O^{-1}$ by its entries, $\tilde{\mathcal{K}}_{ij,\epsilon}^{-1}$ $(i,j = 1,2,3)$, we see that $\mathcal{M}_O = I - \tilde{\mathcal{K}}_O^{-1}\mathcal{K}$ is given by

$$\mathcal{M}_O = \begin{pmatrix} 0 & M_1 & 0 \\ 0 & M_2 & 0 \\ 0 & M_3 & 0 \end{pmatrix} \tag{3.19}$$

where

$$M_i = \tilde{\mathcal{K}}_{i2,\epsilon}^{-1} \mathcal{L}_{uu} \epsilon \mathcal{P}$$

for $i = 1,2,3$. By the iteration (3.7) we have

$$x_O^{n+1} = \tilde{\mathcal{K}}_O^{-1} r + (I - \tilde{\mathcal{K}}_O^{-1} \mathcal{R}_O) x_O^n = \tilde{\mathcal{K}}_O^{-1} r + \mathcal{M}_O x_O^n$$

or by (3.8) with $x_O^n = (v^n, s^n, w^n)^T$

$$\begin{aligned} v^{n+1} &= M_1\, s^n + \tilde{r}_1\,, \\ s^{n+1} &= M_2\, s^n + \tilde{r}_2\,, \\ w^{n+1} &= M_3\, s^n + \tilde{r}_3\,, \end{aligned} \tag{3.20}$$

where $\tilde{r} = \tilde{K}_O^{-1} r = (\tilde{r}_1, \tilde{r}_2, \tilde{r}_3)^T$. It is immediate that the convergence of the sequence $x_O^n = (v^n, s^n, w^n)^T$ is equivalent to the convergence of s^n which proves the first statement of the theorem.

The entry M_2 is given by, see Lemma 3.8,

$$M_2 = \tilde{\mathcal{K}}_{22,\epsilon}^{-1} \mathcal{L}_{uu} \epsilon \mathcal{P} = \mathcal{H}_\epsilon^{-1} \mathcal{L}_{uu} \epsilon \mathcal{P} = I - (\mathcal{H} + \mathcal{L}_{uu}\, \epsilon \mathcal{P})^{-1}\, \mathcal{H}\,.$$

By Theorem 3.6 the condition $\rho(M_2) < 1$ is sufficient for convergence of the sequence s^n and by part 1.) also for the convergence of x_O^n. □

From the result (3.18) it can be seen that we can set the convergence rate of the outer loop by choosing $\epsilon\mathcal{P}$ appropriately. However, the choice of $\epsilon\mathcal{P}$ influences the convergence of the inner loop as well.

3.3.2. Convergence Analysis of the Inner Loop. The convergence properties of the inner loop are described in the following theorem.

THEOREM 3.10. *Under Assumptions* 2.1, 3.2 *and* 3.3 *the following statements hold.*

1.) *The iteration* (3.10) *converges if and only if the iteration*

$$x_2^{k+1} = (I + \epsilon\mathcal{P})^{-1} (\mathcal{L}_{uu}^{-1}\mathcal{H} - I)x_2^k + \bar{r}_2 \tag{3.21}$$

converges, where $\bar{r} = \tilde{K}_I^{-1} r_O$.

2.) *If the spectral radius of the boundary operator* $(I + \epsilon\mathcal{P})^{-1} (\mathcal{L}_{uu}^{-1}\mathcal{H} - I)$ *satisfies*

$$\rho((I + \epsilon\mathcal{P})^{-1} (\mathcal{L}_{uu}^{-1}\mathcal{H} - I)) < 1, \tag{3.22}$$

then the iterates of (3.10) *converge to the solution* x_I^*.

Proof. The inverse of the operator $\mathcal{K}_I$ defined in (3.9) is given by

$$\tilde{\mathcal{K}}_I^{-1} = \begin{pmatrix} h_\phi^{-1} h_u \mathcal{L}_{u,\epsilon}^{-1} h_u^\times h_\phi^{-\times} & -h_\phi^{-1} h_u \mathcal{L}_{u,\epsilon}^{-1} & h_\phi^{-1} \\ -\mathcal{L}_{u,\epsilon}^{-1} h_u^\times h_\phi^{-\times} & \mathcal{L}_{u,\epsilon}^{-1} & 0 \\ h_\phi^{-\times} & 0 & 0 \end{pmatrix}. \tag{3.23}$$

Thus, the operator $\mathcal{M}_I = I - \tilde{\mathcal{K}}_I^{-1}\mathcal{K}$ in (3.3) is for the inner loop explicitly given by

$$\mathcal{M}_I = \begin{pmatrix} h_\phi^{-1} h_u \mathcal{L}_{u,\epsilon}^{-1} (h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi\phi} - \mathcal{L}_{u\phi}) & h_\phi^{-1} h_u \mathcal{L}_{u,\epsilon}^{-1} h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi u} & 0 \\ -\mathcal{L}_{u,\epsilon}^{-1} (h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi\phi} - \mathcal{L}_{u\phi}) & -\mathcal{L}_{u,\epsilon}^{-1} h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi u} & 0 \\ h_\phi^{-\times} \mathcal{L}_{\phi\phi} & h_\phi^{-\times} \mathcal{L}_{\phi u} & 0 \end{pmatrix}. \tag{3.24}$$

Denoting the following composed operators by G, Q, N, C, U,

$$\begin{aligned} G &= -h_\phi^{-1} h_u, \\ Q &= -\mathcal{L}_{u,\epsilon}^{-1} (h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi\phi} - \mathcal{L}_{u\phi}), \\ N &= -\mathcal{L}_{u,\epsilon}^{-1} h_u^\times h_\phi^{-\times} \mathcal{L}_{\phi u}, \\ C &= h_\phi^{-\times} \mathcal{L}_{\phi\phi}, \\ U &= h_\phi^{-\times} \mathcal{L}_{\phi u}, \end{aligned} \tag{3.25}$$

$\mathcal{M}_I$ can be written in the form

$$\mathcal{M}_I = \begin{pmatrix} GQ & GN & 0 \\ Q & N & 0 \\ C & U & 0 \end{pmatrix}. \tag{3.26}$$

Hence the inner iteration with $x_I^k = (x_{I,1}^k, x_{I,2}^k, x_{I,3}^k)^T$ can be written as

$$\begin{aligned} x_{I,1}^{k+1} &= G\,Q\,x_{I,1}^k + G\,N\,x_{I,1}^k + \bar{r}_1, \\ x_{I,2}^{k+1} &= Q\,x_{I,1}^k + N\,x_{I,2}^k + \bar{r}_2, \\ x_{I,3}^{k+1} &= C\,x_{I,1}^k + U\,x_{I,2}^k + \bar{r}_3. \end{aligned} \tag{3.27}$$

Multiply (3.27) by G and substitute the resulting equality into (3.27). This yields

$$x_{I,1}^{k+1} = G\,x_{I,2}^{k+1} + \bar{r}_1 - G\bar{r}_2. \tag{3.28}$$

From this equation it is clear that if $x_{I,2}^k$ converges, then so does $x_{I,1}^k$ and by (3.27) also $x_{I,3}^k$.

To show the second statement, use (3.28) in the form $x_{I,1}^k = G\,x_{I,2}^k + \bar{r}_1 - G\,\bar{r}_2$ to eliminate $x_{I,1}^k$ in (3.27) which gives

$$x_{I,2}^{k+1} = (Q\,G + N)\,x_{I,2}^k + \bar{r}_2 + Q\bar{r}_1 - Q\,G\,\bar{r}_2.$$

By definition (3.25), the operator $Q\,G + N$ coincides with $(I + \epsilon\mathcal{P})^{-1}\,(\mathcal{L}_{uu}^{-1}\mathcal{H} - I)$ (see (2.11)), because

$$\begin{aligned} Q\,G + N = &\ \mathcal{L}_{u,\epsilon}^{-1}\,h_u^{\times}\,h_\phi^{-\times}\,\mathcal{L}_{\phi\phi}\,h_\phi^{-1}\,h_u - \mathcal{L}_{u,\epsilon}^{-1}\,\mathcal{L}_{u\phi}\,h_\phi^{-1}\,h_u - \mathcal{L}_{u,\epsilon}^{-1}\,h_u^{\times}\,h_\phi^{-\times}\,\mathcal{L}_{\phi u} \\ = &\ \mathcal{L}_{u,\epsilon}^{-1}\,(\mathcal{H} - \mathcal{L}_{uu}) \\ = &\ (I + \epsilon\mathcal{P})^{-1}\,(\mathcal{L}_{uu}^{-1}\,\mathcal{H} - I)\,. \end{aligned}$$

Therefore, if $\rho((I + \epsilon\mathcal{P})^{-1}\,(\mathcal{L}_{uu}^{-1}\,\mathcal{H} - I)) = \rho(Q\,G + N) < 1$, then by Theorem 3.6 the iterates $x_{I,2}^k$ converge and by the first assertion also the sequence x_I^k. □

3.3.3. Convergence Analysis in Case of One Inner Iteration. In case only one inner iteration is performed, convergence is determined as follows.

THEOREM 3.11. *Under Assumptions* 2.1, 3.2 *and* 3.3 *the following statements hold.*

1.) *The iteration* (3.13) *converges if and only if the iteration*

$$s^{n+1} = (\mathcal{L}_{u,\epsilon}^{-1}\mathcal{H} - I)\,s^n$$

converges.

2.) *If the spectral radius of the boundary operator* $\mathcal{L}_{u,\epsilon}^{-1}\mathcal{H} - I$ *satisfies*

$$\rho(\mathcal{L}_{u,\epsilon}^{-1}\mathcal{H} - I) < 1, \tag{3.29}$$

then the iterates in (3.13) *converge to the solution* x^* *of* (2.16).

Proof. The inverse of the operator $\tilde{\mathcal{K}} = \tilde{\mathcal{K}}_I$ is given in (3.23). Thus, in this case where only one inner iteration is performed, $\mathcal{M} = I - \tilde{\mathcal{K}}^{-1}\,\mathcal{K}$ is given by

$$\mathcal{M} = \begin{pmatrix} h_\phi^{-1}h_u\mathcal{L}_{u,\epsilon}^{-1}(h_u^{\times}h_\phi^{-\times}\mathcal{L}_{\phi\phi} - \mathcal{L}_{u\phi}) & h_\phi^{-1}h_u\mathcal{L}_{u,\epsilon}^{-1}(h_u^{\times}h_\phi^{-\times}\mathcal{L}_{\phi u} + \mathcal{L}_{uu}\,\epsilon\mathcal{P}) & 0 \\ -\mathcal{L}_{u,\epsilon}^{-1}(h_u^{\times}h_\phi^{-\times}\mathcal{L}_{\phi\phi} - \mathcal{L}_{u\phi}) & -\mathcal{L}_{u,\epsilon}^{-1}(h_u^{\times}h_\phi^{-\times}\mathcal{L}_{\phi u} + \mathcal{L}_{uu}\,\epsilon\mathcal{P}) & 0 \\ h_\phi^{-\times}\mathcal{L}_{\phi\phi} & h_\phi^{-\times}\mathcal{L}_{\phi u} & 0 \end{pmatrix}.$$

We denote the composed operators by G, Q, C, U as before in (3.25) and let N_1 be defined by

$$N_1 \;\; = \;\; -\mathcal{L}_{u,\epsilon}^{-1}\,(h_u^{\times}\,h_\phi^{-\times}\,\mathcal{L}_{\phi u} + \mathcal{L}_{uu}\,\epsilon\mathcal{P})\,.$$

Then $\mathcal{M}$ can be written in the form (3.26) with N replaced by N_1. The proof follows the same lines as for the one of the previous theorem.

By definition of G, Q, C, U in (3.25) and of N_1 above, the operator $Q\,G + N_1$ equals

$$\begin{aligned} &\mathcal{L}_{u,\epsilon}^{-1}\,h_u^{\times}\,h_\phi^{-\times}\,\mathcal{L}_{\phi\phi}\,h_\phi^{-1}\,h_u - \mathcal{L}_{u,\epsilon}^{-1}\,\mathcal{L}_{u\phi}\,h_\phi^{-1}\,h_u - \mathcal{L}_{u,\epsilon}^{-1}\,h_u^{\times}\,h_\phi^{-\times}\,\mathcal{L}_{\phi u} - \mathcal{L}_{u,\epsilon}^{-1}\,\mathcal{L}_{uu}\,\epsilon\mathcal{P} \\ &= Q\,G + N - \mathcal{L}_{u,\epsilon}^{-1}\,\mathcal{L}_{uu}\,\epsilon\mathcal{P} \\ &= \mathcal{L}_{u,\epsilon}^{-1}(\mathcal{H} - \mathcal{L}_{uu}) - \mathcal{L}_{u,\epsilon}^{-1}\mathcal{L}_{uu}\,\epsilon\mathcal{P} \\ &= (I + \epsilon\mathcal{P})^{-1}\,\mathcal{L}_{uu}^{-1}\mathcal{H} - I \\ &= \mathcal{L}_{u,\epsilon}^{-1}\mathcal{H} - I. \end{aligned}$$

□

3.3.4. Discussion of $\mathcal{P}$. The preceding convergence analysis shows that if $\mathcal{P}$ is not present, or, equivalently, if $\epsilon = 0$, there is no outer loop in the nested defect correction process. In that case, the convergence requirements (3.29) and (3.22) coincide and are given by

$$\rho(\mathcal{L}_{uu}^{-1}\mathcal{H} - I) < 1.$$

In general, $\mathcal{L}_{uu}$ will not model the Hessian, $\mathcal{H}$, very well, and convergence is not necessarily ascertained.

In order to solve the system (2.14) efficiently with the defect correction approach, we modify (2.14) through replacing $\mathcal{L}_{uu}$ by $\mathcal{L}_{u,\epsilon} = \mathcal{L}_{uu}(I + \epsilon\mathcal{P})$. We have seen in the preceding Sections 3.3.1, 3.3.2, and 3.3.3, that the convergence rate of the processes can be determined by appropriate choices of $\mathcal{P}$. However, finding $\epsilon\mathcal{P}$ that performs well both in the outer and inner loop requires solving conflicting tasks. Theorem 3.9 suggests that $\mathcal{L}_{uu}\,\epsilon\mathcal{P}$ should be small in the sense that $\mathcal{H} + \mathcal{L}_{uu}\,\epsilon\mathcal{P}$ is only a small perturbation to $\mathcal{H}$. Theorem 3.10 indicates that $\mathcal{L}_{u,\epsilon}$ should approximate $\mathcal{H} - \mathcal{L}_{uu}$ fairly well, because $(I + \epsilon\mathcal{P})^{-1}(\mathcal{L}_{uu}^{-1}\mathcal{H} - I) = \mathcal{L}_{u,\epsilon}^{-1}(\mathcal{H} - \mathcal{L}_{uu})$ is the convergence-governing part. The problem to be solved is

$$\min_{\mathcal{P}} \max\left\{\rho\left(I - (\mathcal{H} + \mathcal{L}_{uu}\,\epsilon\mathcal{P})^{-1}\mathcal{H}\right), \rho\left(I + \epsilon\mathcal{P}\right)^{-1}\left(\mathcal{L}_{uu}^{-1}\mathcal{H} - I\right)\right\}.$$

If only one inner iteration is done, $\mathcal{L}_{u,\epsilon}$ is required to approximate $\mathcal{H}$ as described in (3.29). In both situations, i.e., in the nested defect correction and in the case with only one inner iteration, some knowledge of the operators involved is necessary for an appropriate choice of $\mathcal{P}$. However, the defect correction process with one inner iteration is easier to apply than the nested defect correction process (because it allows for a closed representation), and thus preferable whenever applicable. If there is not much information available on the Hessian, choosing a "small" $\epsilon\mathcal{P}$ and applying nested defect correction seems to be more promising. We will discuss the choice of $\mathcal{P}$ for our example problem in Sections 4.5 and 4.7.

4. Application: The Small Perturbation Potential Problem. We consider an optimal control problem governed by a partial differential equation. The example problem is motivated by problems of aerodynamic optimization. In the following, we derive the example problem and state the equations relevant to the approach delineated in Sections 2 and 3. Subsequently we turn to the discretization and the finite–dimensional solution of the optimization problem. We include a convergence estimate for the example using local mode analysis.

4.1. The Small Perturbation Potential Equation. We start with a short derivation of the state equation that governs our optimal control problem. We assume inviscid irrotational flow modeled by the continuity equation,

$$\nabla(\rho\vec{v}) = 0,$$

where ρ is the density of the fluid and $\vec{v}$ the velocity field. The potential function, Φ, is defined with the relation

$$\vec{v} = \nabla\Phi.$$

The density is related to the potential by the isentropic density law (see, e.g., [13]). We consider a slender body aligned with the x–axis, and define a perturbation potential ϕ by

$$\Phi = U_\infty(x + \phi),$$

where U_∞ is the free–stream velocity. Under the assumption of a dominating x–component of the velocity field, and neglecting terms that are proportional to ϕ_x^2 and to ϕ_y^2, the following first-order transonic small-perturbation is obtained (Prandtl-Glauert equation):

$$(1 - M_\infty^2)\phi_{xx} + \phi_{yy} = 0.$$

If $y = u(x)$ is the equation of the surface of the slender body, it is possible to set the surface boundary condition on the x–axis, that is at $y = 0$. The presence of the slender body will appear in the computation only through the boundary condition

$$v = (U_\infty + u)u_x \approx U_\infty u_x. \tag{4.1}$$

In terms of the perturbation potential, (4.1) is given by the normal derivative

$$\phi_n = u_x.$$

The boundary conditions at the far–field for the perturbation potential are set such that it does not affect the far–field velocity.

The small perturbation potential equation has been for a long time the basis for potential flow theories as it is a simplified form valid for flow fields along slender bodies aligned with the x–axis. We turn to form a simple optimal control problem based on that model.

4.2. The Optimal Control Problem. The small-perturbation potential problem allows us to study a shape optimization problem with a boundary control model defined on a fixed domain, thus avoiding the complication of a changing geometry.

We consider the following minimization problem,

$$\min_{(\phi,u)} \quad \frac{1}{2}\int_\Gamma (\phi_x - \phi_x^d)^2\, ds + \frac{\eta_1}{2}\int_\Gamma u^2\, ds + \eta_2 \int_\Gamma \phi_x u\, ds\,, \tag{4.2}$$

subject to the following state equations,

$$\begin{aligned} (1-M_\infty^2)\phi_{xx} + \phi_{yy} &= 0 \quad && \text{in } \Omega = (0,1)^2, \\ \phi_n &= u_x \quad && \text{on } \Gamma = \{(x,0) : 0 < x < 1\}, \\ \phi_n &= 0 \quad && \text{on } \Gamma_l \cup \Gamma_r\,, \\ \phi &= 0 \quad && \text{on } \Gamma_t. \end{aligned} \tag{4.3}$$

Here, the parts of the boundary Γ_l, Γ_r, and Γ_t are given by $\Gamma_l = \{(0,y) : 0 < y < 1\}$, $\Gamma_r = \{(1,y) : 0 < y < 1\}$, and $\Gamma_t = \{(x,1) : 0 < x < 1\}$.

We now give a short explanation of the different terms in the cost functional (4.2). The first term is proportional to a pressure matching term since in the small disturbance model the pressure of the flow on the slender body is proportional to the derivative of the potential in the flow direction, ϕ_x. The desired "pressure distribution", ϕ_x^d, is given. The second term is a penalty on the control and η_1 is a parameter. The third term is artificially introduced to the objective function to better model the structure of problems in applications since in aerodynamic optimization problems often non–zero terms $\mathcal{L}_{u\phi}$ and $\mathcal{L}_{\phi u}$ are present.

4.3. Existence and Uniqueness of an Optimal Control. Let us assume that the design, $u(x)$, belongs to the subspace of functions that can be spanned by the basis $\{\sin(2\pi kx)\}_{k=1}^n$, i.e.,

$$u(x) = \sum_{k=1}^{n} \hat{u}_k \sin(2\pi kx), \tag{4.4}$$

where $\{\hat{u}_k\}_{k=1}^n$ are real numbers and n is a positive integer. For that choice of $u(x)$ there exists a solution $\phi(x,y)$ of the minimization problem (4.2) subject to the state equation (4.3) of the form

$$\phi(x,y) = \sum_{k=1}^{n} \hat{u}_k \psi_k(y) \cos(2\pi kx)\,, \tag{4.5}$$

where then functions $\psi_k(y)$ are given by

$$\psi_k(y) = \frac{1}{\gamma}\left(-\tanh(2\pi\gamma k)\cosh(2\pi\gamma k y) + \sinh(2\pi\gamma k y)\right) \tag{4.6}$$

for $\gamma = \sqrt{1 - M_\infty^2}$. By inspection the above ϕ satisfies the state equation (4.3).

REMARK 4.1. *By the solution* (4.5) *the operator* h_ϕ *has the symbol* $(2\pi k)/\psi_k(y)$. *(For the definition of the symbol of a differential operator see, e.g., [21], p.38.) Since the Fourier transformation is a homeomorphism we conclude that the operator* h_ϕ *satisfies Assumption 2.1.*

THEOREM 4.2. *There exists a unique solution* $u^*(x)$ *to the optimal control problem* (4.2) *subject to the state equation* (4.3).

Proof. Let us denote the vector of design coefficients by u_N: $u_N^T = (\hat{u}_1, \cdots, \hat{u}_n)$. A direct substitution of the solution (4.5) into the cost functional (4.2) results in a leading quadratic term of the form $u_N^T Q u_N$ with Q being a positive definite matrix. This proves that the minimization problem has a unique solution. □

4.4. Optimality Conditions and the Newton Step. The first–order optimality conditions of the problem (2.1), as outlined in Section 2.2, are the state, costate, and design equations (2.5), (2.6) (2.7). The state equation is given in (4.3). The adjoint equation for this problem takes the form

$$
\begin{aligned}
(1 - M_\infty^2)\lambda_{xx} + \lambda_{yy} &= 0 && \text{in } \Omega, \\
\lambda_n &= -(\phi_{xx} - \phi_{xx}^d) + \eta_2\, u_x && \text{on } \Gamma, \\
\lambda_n &= 0 && \text{on } \partial\Omega - \Gamma.
\end{aligned} \tag{4.7}
$$

From the design equation (2.7) we get the gradient

$$
g = -\lambda_x + \eta_1\, u + \eta_2\, \phi_x \qquad \text{on } \Gamma. \tag{4.8}
$$

The Newton step satisfies (2.14) with the operators

$$
\mathcal{L}_{\phi\phi} = \left\{ \begin{array}{r} 0|_\Omega \\ -\partial_{xx}|_\Gamma \end{array} \right\}, \quad \mathcal{L}_{uu} = \eta_1 \cdot I|_\Gamma, \quad \mathcal{L}_{u\phi} = \mathcal{L}_{\phi u} = \left\{ \begin{array}{r} 0|_\Omega \\ \eta_2 \partial_x|_\Gamma \end{array} \right\}
$$

and

$$
h_\phi = h_\phi^\times = \left\{ \begin{array}{r} (1 - M_\infty^2)\partial_{xx} + \partial_{yy}|_\Omega \\ -\partial_n|_\Gamma \end{array} \right\}, \quad h_u = -h_u^\times = \left\{ \begin{array}{r} 0|_\Omega \\ -\partial_x|_\Gamma \end{array} \right\}.
$$

Explicitly, the Newton step, s, satisfies the following system of PDEs:

$$
\begin{aligned}
(1 - M_\infty^2)\, w_{xx} + w_{yy} &= 0 && \text{in } \Omega, \\
w_n - v_{xx} + \eta_2\, s_x &= 0 && \text{on } \Gamma.
\end{aligned} \tag{4.9}
$$

$$
\begin{aligned}
\eta_1\, s + \eta_2\, v_x - w_x &= -g && (x, 0) \text{ on } \Gamma, \\
w(1) &= g(1).
\end{aligned} \tag{4.10}
$$

$$
\begin{aligned}
(1 - M_\infty^2)\, v_{xx} + v_{yy} &= 0 && (x, y) \text{ in } \Omega, \\
v_n - s_x &= 0 && (x, 0) \text{ on } \Gamma, \\
s(0) &= 0.
\end{aligned} \tag{4.11}
$$

The defect correction process described in Section 3 will now be applied to the example problem introduced in Section 4.2. Convergence of the solution process (3.4) is governed by the singular values of the operators $\mathcal{M}_O$, $\mathcal{M}_I$, and $\mathcal{M}$ derived in Sections 3.3.1, 3.3.2, and 3.3.3. The operators for the small perturbation potential problem are given in Section 4.4 so that $\mathcal{M}_O$, $\mathcal{M}_I$, and $\mathcal{M}$ are easily found for the example.

Knowledge of the operators allows to choose $\mathcal{P}$ such that small convergence rates are obtained. We now use local mode analysis of the convergence–governing operator to choose the operator $\mathcal{P}$. Similar analysis has been introduced in the past to approximate the reduced Hessian of optimization problems governed by PDEs (see, e.g., [1], [2], [3]).

4.5. Choice of $\mathcal{P}$ by local mode analysis of the PDEs. The local mode analysis is performed locally around a point on the boundary Γ, ignoring the boundary conditions on $\partial\Omega - \Gamma$. Thus for a boundary value problem, as we have, it is only an approximation. We deliberately do not insist on the exact analysis since it can not be done in general applications while the given analysis can be applied (after linearization and freezing of coefficients when the problem is nonlinear).

We choose to use the splitting of Section 3.2.3, i.e., the case of one inner iteration. We have seen in that section that convergence depends on the eigenvalues of the operator $\mathcal{T} = \mathcal{L}_{u,\epsilon}^{-1}\mathcal{H} - I$. We study one Fourier component of the error,

$$e(x,y) = \hat{e}(\omega_1,\omega_2)\, e^{\mathbf{i}(\omega_1 x + \omega_2 y)}\,.$$

The interior equation in (4.3) relates ω_1 and ω_2 by

$$(1 - M_\infty^2)\omega_1^2 + \omega_2^2 = 0. \tag{4.12}$$

We choose the decaying mode solution for Equation (4.12),

$$\omega_2 = i\sqrt{1 - M_\infty^2}\,|\omega_1|. \tag{4.13}$$

The boundary equation implies that

$$\sqrt{1 - M_\infty^2}\,|\omega_1|\,\hat{\phi} = i\omega_1\,\hat{u}.$$

We arrive at the Fourier symbols of the operators in the convergence–governing operator $\mathcal{T}$:

$$\begin{aligned}
\hat{h}_\phi &= \hat{h}_\phi^\times &&= \sqrt{1 - M_\infty^2}\,|\omega_1|\,,\\
\hat{h}_u &= -\hat{h}_u^\times &&= -i\omega_1\,,\\
\hat{\mathcal{L}}_{\phi\phi} &= -\hat{D}_{xx} &&= \omega_1^2\,,\\
\hat{\mathcal{L}}_{\phi u} &= -\hat{\mathcal{L}}_{u\phi} &&= -\eta_2\, i\omega_1\,,\\
\hat{\mathcal{L}}_{uu} &= \eta_1\,. &&
\end{aligned} \tag{4.14}$$

These imply that the Fourier symbol of the Hessian is given by

$$\hat{\mathcal{H}} = \hat{h}_u^\times\,\hat{h}_\phi^{-\times}\,\hat{\mathcal{L}}_{\phi\phi}\,\hat{h}_\phi^{-1}\,\hat{h}_u - \hat{h}_u^\times\,\hat{h}_\phi^{-\times}\,\hat{\mathcal{L}}_{\phi u} - \hat{\mathcal{L}}_{u\phi}\,\hat{h}_\phi^{-1}\,\hat{h}_u + \hat{\mathcal{L}}_{uu} = \omega_1^2 - 2\eta_2\frac{\omega_1^2}{|\omega|} + \eta_1.$$

By (3.29), the choice of the operator $\mathcal{P}$ is such that $\mathcal{P} \approx \mathcal{L}_{uu}^{-1}\mathcal{H}$. We obtain the approximated symbol of the desired operator $\mathcal{P}$ as

$$\hat{\mathcal{P}} = \frac{1}{\eta_1}\Big(\omega_1^2 - 2\eta_2\frac{\omega_1^2}{|\omega_1|}\Big)\,. \tag{4.15}$$

Since the second term in (4.15) does not correspond to a differential operator, and because it is of lower order than the first term, our first approximation of $\mathcal{P}$ is given by

$$\mathcal{P} = -\frac{1}{\eta_1}D_{xx}. \tag{4.16}$$

We now turn to use the same local mode analysis to approximate the asymptotic convergence rate of the defect correction process for low mesh size h. The above symbols (4.14) imply that the symbol of the convergence–governing operator $\mathcal{T}$ is, where $\epsilon\hat{\mathcal{P}}$ is taken to be $\frac{1}{\eta_1}\omega_1^2$,

$$\hat{\mathcal{T}} = \hat{\mathcal{L}}_{u,\epsilon}^{-1}\hat{\mathcal{H}} - 1 = (1 + \epsilon\hat{\mathcal{P}})^{-1}\hat{\mathcal{L}}_{uu}^{-1}\hat{\mathcal{H}} - 1.$$

Using Parseval identity we estimate an upper bound for the convergence rate of the iterates of the defect correction process by

$$\mu \leq \max_{\omega} \left(\hat{T}^*(\omega)\, \hat{T}(\omega) \right)^{\frac{1}{2}}, \tag{4.17}$$

where $\omega = \frac{k\pi}{n}$ for k ranging from $1, \ldots, n$. Here, n is the number of grid points on the boundary Γ, and $\hat{T}^*$ is the complex conjugate of $\hat{T}$.

4.6. The Discretization. We define a uniform grid on the domain Ω, containing m grid points. The perturbation potential, ϕ, is defined on the grid vertices, and the control is defined on the mid–interval points on the boundary Γ, i.e., $\phi \in I\!R^m$ and $u \in I\!R^n$, with $n = \sqrt{m} - 1$. We then apply a second order finite difference discretization to the problem (4.2), (4.3). The stencils can be found in (4.24) and (4.26). The resulting finite dimensional problem is to minimize a quadratic functional under linear constraints,

$$\min_{(\phi_N, u_N)} F(\phi_N, u_N) \text{ s.t. } A\phi_N + Bu_N = b, \tag{4.18}$$

where the discretized objective function can be written as

$$F = \frac{1}{2}\, \phi_N^T H_{\phi\phi}\, \phi_N + \frac{\eta_1}{2}\, u_N^T H_{uu}\, u_N + \eta_2\, \phi_N^T H_{\phi u}\, u_N + \phi_N^T c + u_N^T d. \tag{4.19}$$

The discrete Lagrangian is given by

$$L(\phi_N, u_N, \lambda_N) = F + \lambda_N^T (A\phi_N + Bu_N - b). \tag{4.20}$$

Note that for this (quadratic) problem the second partial derivatives $L_{\phi\phi}$, $L_{\phi u}$ and L_{uu} of the Lagrangian coincide with $H_{\phi\phi}$, $H_{\phi u}$ and H_{uu}, respectively.

Applying the first order optimality conditions (Karush–Kuhn–Tucker conditions, see, e.g., [20]), we have to solve the system

$$\begin{pmatrix} L_{\phi\phi} & \eta_2\, L_{\phi u} & A^T \\ \eta_2\, L_{u\phi} & \eta_1\, L_{uu} & B^T \\ A & B & 0 \end{pmatrix} \begin{pmatrix} \phi_N \\ u_N \\ \lambda_N \end{pmatrix} = \begin{pmatrix} -c \\ -d \\ b \end{pmatrix},$$

where

$$\phi_N \in I\!R^m,\ u_N \in I\!R^n,\ \lambda_N \in I\!R^m,\ c \in I\!R^m,\ d \in I\!R^n,\ b \in I\!R^m,$$
$$L_{\phi\phi} \in I\!R^{m\times m},\ L_{uu} \in I\!R^{n\times n},\ L_{\phi u} \in I\!R^{m\times n},\ A \in I\!R^{m\times m},\ B \in I\!R^{m\times n}.$$

If A is invertible, the discrete state equation $A\phi_N + Bu_N = b$ can be solved for ϕ_N, $\phi_N(u_N) = A^{-1}(b - Bu_N)$. Thus, we can define the discrete unconstrained problem

$$\min_{u_N} J(u_N) = F(\phi_N(u_N), u_N)\,. \tag{4.21}$$

In order to get the gradient, g_N, of the unconstrained problem (4.21), the state and the costate equations must be solved exactly. This means that for a given control u_N the following set of equations has to be satisfied,

$$\begin{pmatrix} L_{\phi\phi} & A^T \\ A & 0 \end{pmatrix} \begin{pmatrix} \phi_N \\ \lambda_N \end{pmatrix} = \begin{pmatrix} -c - L_{\phi u}\, u_N \\ b - Bu_N \end{pmatrix}.$$

Then the gradient can be computed as

$$g_N = B^T \lambda_N + \eta_1\, L_{uu} u_N + \eta_2\, L_{\phi u} \phi_N + d.$$

The Hessian of the discrete unconstrained problem (4.21) is given by

$$H = B^T A^{-T} L_{\phi\phi} A^{-1} B - L_{u\phi} A^{-1} B - B^T A^{-T} L_{\phi u} + L_{uu}. \tag{4.22}$$

The Newton step, s_N, of the discretized unconstrained problem (4.21) can be computed by solving the system

$$\begin{pmatrix} L_{\phi\phi} & L_{\phi u} & A^T \\ L_{u\phi} & L_{uu} & B^T \\ A & B & 0 \end{pmatrix} \begin{pmatrix} v_N \\ s_N \\ w_N \end{pmatrix} = \begin{pmatrix} 0 \\ -g_N \\ 0 \end{pmatrix}. \tag{4.23}$$

4.7. Choice of $\mathcal{P}$ by local mode analysis of the Finite Difference Equations. We now perform the local mode analysis, similar to the local mode analysis in Section 4.5, taking into account the specific discretization. By analyzing the finite difference equations we hope to get a better approximation of the reduced Hessian and as a result a better approximation of the operator $\mathcal{P}$. Note that although the operators analysed in this section are finite difference operators and not differential operators, we still denote them as before to avoid excessive notation.

The discrete interior equation has the form

$$a\phi_{i,j} + b(\phi_{i,j+1} + \phi_{i,j-1}) + c(\phi_{i+1,j} + \phi_{i-1,j}) = 0 \tag{4.24}$$

where

$$b = \frac{1}{h^2}, \quad c = (1 - M_\infty^2)b, \quad a = (1 - M_\infty^2)(-2b) - 2b.$$

We study one Fourier component of the error,

$$e(x, y) = \hat{e}(\theta_1, \theta_2)\, e^{i(\theta_1 \frac{x}{h} + \theta_2 \frac{y}{h})},$$

where the mesh sizes in the x– and in the y–directions are assumed to be a constant h. The discrete interior equation (4.24) relates θ_1 and θ_2 ($a + 2(c\cos\theta_1 + b\cos\theta_2) = 0$) to

$$\theta_2 = \cos^{-1}[-\frac{a + 2c\cos\theta_1}{2b}]. \tag{4.25}$$

The discrete boundary equation has the form

$$a\phi_{i,j} + 2b\phi_{i,j+1} + c(\phi_{i+1,j} + \phi_{i-1,j}) = \frac{1}{h}(u_i - u_{i-1}). \tag{4.26}$$

In terms of the Fourier component of the error in ϕ, this implies

$$(a + 2c\cos\theta_1 + 2be^{i\theta_2})\hat{\phi}(\theta_1, \theta_2) = \frac{2i}{h}\, \sin(\frac{\theta_1}{2})\, \hat{u}(\theta_1), \tag{4.27}$$

or equivalently,

$$2bi\sin\theta_2\, \hat{\phi} = \frac{2i}{h}\, \sin(\frac{\theta_1}{2})\, \hat{u}. \tag{4.28}$$

Using the identity $\sin\cos^{-1} x = \sqrt{1 - x^2}$ together with Equations (4.25), (4.27), and (4.28), we arrive at the symbol of the operator h_ϕ, given by

$$\hat{h}_\phi = \hat{h}_\phi^\times = \frac{i}{h}\sqrt{1 - \Big(\frac{a + 2c\cos\theta_1}{2b}\Big)^2}.$$

The other operators in the convergence–governing operator, $L_{u,\epsilon}^{-1}H - I$, have the following symbols:

$$
\begin{aligned}
\hat{h}_u &= -\hat{h}_u^{\times} = -\tfrac{2i}{h}\,\sin(\tfrac{\theta_1}{2}) = -\tfrac{i}{h}\,\sqrt{2(1-\cos\theta_1)}\,, \\
\hat{\mathcal{L}}_{\phi\phi} &= -\hat{D}_{xx} = \tfrac{2}{h^2}(1-\cos\theta_1)\,, \\
\hat{\mathcal{L}}_{\phi u} &= -\hat{\mathcal{L}}_{u\phi} = \eta_2\,\hat{h}_u\,, \\
\hat{\mathcal{L}}_{uu} &= \eta_1 .
\end{aligned}
\tag{4.29}
$$

The reduced Hessian (2.11) contains the operator $h_\phi^{-1}h_u$ and its adjoint. The expression $\hat{h}_\phi^{-1}\hat{h}_u$ can be simplified to

$$
\hat{h}_\phi^{-1}\hat{h}_u = \frac{1}{\sqrt{-(1-M_\infty^2) - \frac{1}{2}(1-M_\infty^2)^2(1-\cos\theta_1)}}\,.
$$

This leads to our second choice of the operator $\mathcal{P}$,

$$
\mathcal{P} = -\frac{1}{\eta_1}\left((1-M_\infty^2)I + \frac{h^2}{4}(1-M_\infty^2)^2 D_{xx}\right)^{-1} D_{xx}\,. \tag{4.30}
$$

Having chosen the operator $\mathcal{P}$, the asymptotic convergence estimates follows as in Section 4.5.

4.8. The Defect Correction Process. On the discrete level, the linear system (4.23), which we denote by

$$
K\,x_N = r_N\,, \tag{4.31}
$$

has to be solved in order to compute the Newton step, s_N. The defect correction process is described by the iteration

$$
\tilde{K}\,x_N^{+} = r_N - R\,x_N^{c}\,, \tag{4.32}
$$

where $\tilde{K}$ and R define a splitting of K, i.e., $\tilde{K} + R = K$. Convergence of the solution process (4.32) is governed by the singular values of the operators $\mathcal{M}_O$, $\mathcal{M}_I$, and $\mathcal{M}$ derived in Sections 3.3.1, 3.3.2, and 3.3.3. Their discrete counterparts M_O, M_I, and M are analyzed with local mode analysis in Section 4.7. Convergence of the outer loop depends on the spectrum $\rho(M_O) = \rho(I - (H + L_{uu}\,\epsilon P)^{-1}\,H)$. Convergence in the inner loop is governed by $\rho(M_I) = \rho((I + \epsilon P)^{-1}\,(L_{uu}^{-1}H - I))$. If only one inner iteration is done, we investigate $\rho(M) = \rho(L_{u,\epsilon}^{-1}H - I)$. One specific choice of $\mathcal{P}$ was given in Section 4.5, motivated by the local mode analysis of the differential operators. The corresponding matrix is

$$
P_1 = -\frac{1}{1-M_\infty^2}\,D_{xx,N}\,. \tag{4.33}
$$

Analyzing further the finite difference equations, the local mode analysis in Section 4.7 motivates secondly to use the matrix

$$
P_2 = -\left(\frac{1}{\eta_1}(1-M_\infty^2)I + \frac{h^2}{4}(1-M_\infty^2)^2 D_{xx,N}\right)^{-1} D_{xx,N}\,. \tag{4.34}
$$

The main computational work in the defect correction process with the splittings defined in Sections 3.3.1, 3.3.2, and 3.3.3 is the solve with h_ϕ and its adjoint. This is the solution of the linearized state equation and of the adjoint equation, which amounts to the solution of two linear PDEs in our example. On the discrete level, h_ϕ is represented by A, and the main computational work in each iteration is the solve with A and A^T.

4.9. A Preconditioned Krylov Subspace Method. Linear systems like the above (4.31) can be solved with Krylov subspace methods, e.g., the well–known Krylov subspace method for general matrices GMRES, see [23]. However, with ill–conditioned problems, as the one given in Equations (4.2) and (4.3), the number of steps these methods require can be as high as the dimension of the linear system, if they do not fail altogether. A high number of steps usually presents an unacceptable computational effort. However, Krylov subspace methods can be very fast and efficient for well–conditioned systems, cf. [25], [29]. Under certain assumptions, see [6], superlinear convergence can be proven for GMRES. In the following we will see that the results furnished by the local mode analysis for the modified system defect correction can enhance the performance of preconditioned GMRES iterations.

The preconditioner we use is closely related to the splittings we propose for the defect correction; it is in fact identical to the splitting matrix $\tilde{K}$ introduced in Section 3.2.3. Thus, the linear system (4.31) is replaced by the preconditioned system

$$\tilde{K}^{-1} K x_N = \tilde{K}^{-1} r_N . \tag{4.35}$$

In each iteration of the preconditioned GMRES, the matrix–vector–product $\tilde{K}^{-1} K z = z^+$ must be computed. This can be done successively by solving three linear subsystems, in a similar way as described in Section 3.2.4. In this respect, the work required in one GMRES iteration is roughly the same as in one defect correction iteration, i.e., one solve with A and one solve with A^T (see [4]). However, the implementation of GMRES is more difficult than that of the defect correction process. For example, re–orthogonalization, which can be very costly as well, is often necessary. In addition, storage requirements increase as the iteration progresses, thus rendering the method unattractive of very large problems. For these issues see, e.g., [23], [14].

The eigenvalues of the preconditioned system matrix $\tilde{K}^{-1} K$ are bounded below in absolute value by 1. The number of eigenvalues distinct from 1 are at most n, where n is the number of design variables. For these results see [4]. Since the performance of GMRES, similar to that of other Krylov subspace methods, depends on the eigenvalue distribution of the underlying system matrix (see, e.g., [6]), the theory indicates that GMRES will take not more than $n + 1$ steps. The numerical results are described in the following Section 4.10.

4.10. Numerical Results. In the numerical tests, we did not restrict the design variable, $u(x)$, to the subspace of sin functions, thus extending the computations beyond the theoretical treatment in Section 4.3.

We show results of the defect correction process for the case of one inner iteration as described in Section 3.2.3. The system is modified with P_1 and P_2 defined in (4.33) and (4.34), respectively. For the parameter M_∞, the values $0.0, 0.1, 0.5, 0.9$ are used in the computations. We consider the combination of cost function parameters $\eta_1 = 1.0$, $\eta_2 = 0.0$, and $\eta_1 = 1.0$, $\eta_2 = -1.0$. For the same combinations of parameters, GMRES is tested on the preconditioned system $\tilde{K}^{-1}K$. We do not give the convergence history for unpreconditioned GMRES, but state that the number of iterations required for convergence is almost equal to the dimension $2m + n$ of the system in all considered cases.

In Figures 5.1 and 5.2, typical convergence behavior of the defect correction process and preconditioned GMRES can be seen. The figures depict the main results, the small and mesh–independent convergence rates of the defect correction process that are exhibited right from the beginning of the iterations, and the improvements in GMRES achieved with the suggested preconditioning.

In the tables, results are shown for systems of dimension 54, which corresponds to a 4×4 grid, up to dimension 8514, which is a grid of 64×64 points. We always chose the discretization in y–direction equal to the discretization in x–direction and refer to this number by the dimension n of the design space. The number m of state variables is given as $m = (n + 1)^2$.

Stopping criterion for all iterations is a threshold 10^{-5} for the l_2–norm of the residual, i.e., we stop if the following

requirement is met,

$$\|K\,x_N - r_N\|_{l_2} = \sqrt{\frac{1}{2m+n}\sum_{i+1}^{2m+n}(K\,x_N - r_N)_i^2} < 10^{-5}.$$

Performance of the defect correction process is shown in Tables 5.3 and 5.4. In the considered cases we only allow for one inner iteration. For the choice of cost function parameters $\eta_1 = 1.0$, $\eta_2 = 0.0$ in Table 5.3, the terms $L_{\phi u}$, $L_{u\phi}$ in K vanish. This does not only simplify the convergence analysis, but also often admits a faster numerical solution than the second choice of nonzero η_2. This is easily seen by a comparison with Table 5.4. The dimension of the design space, n, and of the entire system are given together with the numerical results for P_1 and P_2 defined in (4.33) and (4.34), respectively. For both choices of P, the number of steps until solution and the CPU required by the iterative process are given. The convergence rate, the ratio of successive errors, is the asymptotic convergence rate valid at the end of the defect correction iterations. This rate is approximately a constant throughout the defect correction process for given parameters and grid. The largest eigenvalue of the matrix M defined by the splitting is a close upper bound for the convergence rate. The computations are done for four different parameters M_∞, $M_\infty = 0.0, 0.1, 0.5, 0.9$. Although the original system becomes increasingly ill–conditioned as M_∞ approaches 1, the modified system defect correction still performs well for $M_\infty = 0.9$. The convergence of this defect correction is mesh–independent. The asymptotic convergence rates, in the limit of mesh–size going to zero, furnished by the local mode analysis for each specific combination of parameters is given in Table 5.5. The discrepancy between the results of the local mode analysis and the actual convergence rates of the defect correction process is due to the fact that the considered domain is finite, which is not taken into account by the local mode analysis performed here.

Performance of the preconditioned GMRES is shown in Tables 5.2 and 5.1. Again, the dimensions of the design space and of the entire system are given together with the number of steps until solution and the required CPU. It can be seen that the required CPU times for iterations of preconditioned GMRES and of the defect correction process are of the same order of magnitude. Since the convergence rate of GMRES is in general not constant throughout the iteration, it is not considered in the tables. Qualitatively, the convergence rate is depicted in Figure 5.1.

In our computations with GMRES we have not only considered P_1 and P_2 given in (4.33) and (4.34) , but also $P = 0$, i.e., the preconditioner $\tilde{K}$ with the entry $L_{u,0} = L_{uu}$. It can be seen that the number of steps with this preconditioner roughly equals $n/2$ (Table 5.2, $L_{\phi u} = 0$, $L_{u\phi} = 0$) or n (Table 5.1, $L_{\phi u} \neq 0$, $L_{u\phi} \neq 0$), respectively. Investing the computational effort of introducing P_1 or P_2 as suggested by the local mode analysis pays out in a low and mesh–independent number of iterations.

All computations were done with Matlab on a SUN 2 X UltraSPARC-II with 2Gb RAM.

5. Discussion and Concluding Remarks. We propose a modified defect correction process to solve efficiently the (KKT–)system of equations composed of the necessary optimality conditions for optimization problems governed by state equations. The new method is simple to apply and embed into existing codes. It requires to solve successively the linearized state and the costate (adjoint) equations with different right hand sides in each iteration. These linear equations are obtained by first modifying the KKT system, $\mathcal{K}$, with the introduction of a "preconditioning" operator $\mathcal{P}$, and then splitting the system into two parts, $\mathcal{K} = \tilde{\mathcal{K}}(\mathcal{P}) + \mathcal{R}(\mathcal{P})$. The solution is obtained by a defect correction process that requires one solve with the operator $\tilde{\mathcal{K}}(\mathcal{P})$ in each iteration. Convergence theory is provided in the paper for the different splittings that we propose. We also suggest to use the operator $\tilde{\mathcal{K}}(\mathcal{P})$ as a preconditioner for GMRES. The introduction of the operator $\mathcal{P}$ is crucial for fast convergence. We advocate to use the structure induced by the governing PDE. This can be done by a local mode analysis of either the PDE, or of its discretized equations. We test both the defect correction process and the preconditioned GMRES on a model problem that mimics an aerodynamic shape optimization problem. We obtain for both methods fast and mesh–independent convergence.

The numerical results are consistent both with the convergence theory in the paper and with the approximated local mode analysis for the asymptotic convergence rate. An extension of the approach to optimization problems with inequality constraints will be treated elsewhere. Another application might be the area of multidisciplinary design and optimization problems. When considering a large system of equations, obtained for example in those problems, a splitting of the KKT system can be applied twice, once to decouple the large multidisciplinary KKT system into subsystems [1], and a second time to solve efficiently each of the subsystems. Such a method might require the introduction of an operator $\mathcal{P}$ for the different subsystems as well as for the large system.

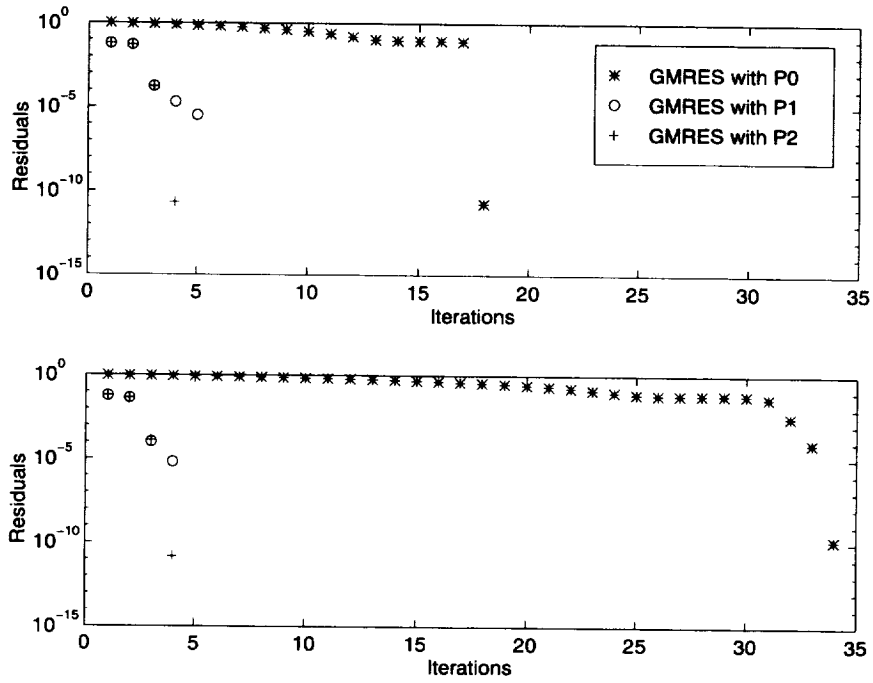


FIG. 5.1. *Residual development of preconditioned GMRES for grid sizes* $n = 32$ *(upper figure) and* $n = 64$ *(lower figure). (Problem parameters:* $\eta_1 = 1.0$, $\eta_2 = 0.0$, $M_\infty = 0.5$*)*

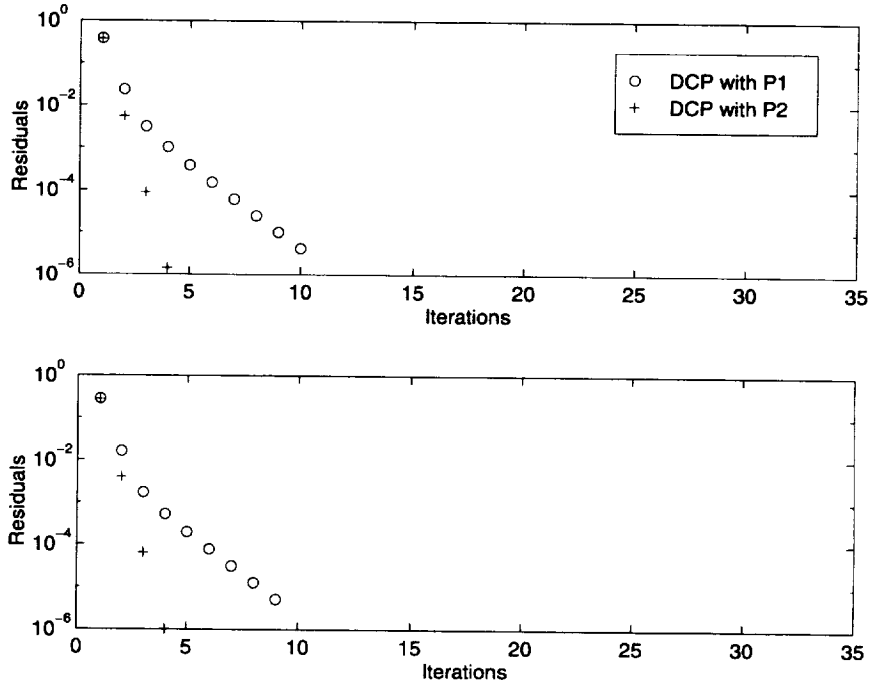


FIG. 5.2. *Residual development of defect correction process for grid sizes* $n = 32$ *(upper figure) and* $n = 64$ *(lower figure). (Problem parameters:* $\eta_1 = 1.0$, $\eta_2 = 0.0$, $M_\infty = 0.5$*)*

TABLE 5.1

Performance of preconditioned GMRES for cost function parameters $\eta_1 = 1.0, \eta_2 = -1.0$.

$(M_\infty = 0.0)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	6	2.00e-01	6	9.00e-02	6	9.00e-02
8	170	10	1.80e-01	6	9.00e-02	7	1.00e-01
16	594	18	8.20e-01	6	4.10e-01	7	4.40e-01
32	2210	32	6.60e+00	7	3.10e+00	6	3.12e+00
64	8514	57	8.26e+01	7	3.29e+01	6	3.17e+01

$(M_\infty = 0.1)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	6	8.00e-02	6	4.00e-02	6	4.00e-02
8	170	10	1.70e-01	6	9.00e-02	7	1.00e-01
16	594	18	7.50e-01	6	3.80e-01	7	4.00e-01
32	2210	32	6.51e+00	7	2.64e+00	6	2.72e+00
64	8514	57	8.36e+01	7	2.46e+01	6	2.48e+01

$(M_\infty = 0.5)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	6	8.00e-02	6	5.00e-02	6	5.00e-02
8	170	10	1.70e-01	6	8.00e-02	6	1.00e-01
16	594	18	7.50e-01	6	3.70e-01	6	4.00e-01
32	2210	33	6.71e+00	6	2.63e+00	6	2.74e+00
64	8514	59	8.68e+01	6	2.38e+01	6	2.52e+01

$(M_\infty = 0.9)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	6	9.00e-02	5	5.00e-02	5	5.00e-02
8	170	10	1.60e-01	6	8.00e-02	6	9.00e-02
16	594	18	7.20e-01	6	3.20e-01	6	3.70e-01
32	2210	34	6.75e+00	6	2.48e+00	6	2.54e+00
64	8514	66	9.37e+01	6	2.32e+01	6	2.43e+01

TABLE 5.2

Performance of preconditioned GMRES for cost function parameters $\eta_1 = 1.0, \eta_2 = 0.0$.

$(M_\infty = 0.0)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	4	8.00e-02	4	4.00e-02	4	3.00e-02
8	170	6	1.30e-01	6	8.00e-02	4	6.00e-02
16	594	10	5.30e-01	6	3.40e-01	4	3.00e-01
32	2210	18	4.19e+00	5	2.55e+00	4	2.18e+00
64	8514	33	5.30e+01	4	2.34e+01	4	2.15e+01

$(M_\infty = 0.1)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	4	8.00e-02	4	4.00e-02	4	3.00e-02
8	170	6	1.30e-01	6	8.00e-02	4	6.00e-02
16	594	10	5.30e-01	6	3.50e-01	4	2.90e-01
32	2210	18	4.23e+00	5	2.49e+00	4	2.20e+00
64	8514	33	5.31e+01	4	2.35e+01	4	2.14e+01

$(M_\infty = 0.5)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	4	8.00e-02	4	4.00e-02	4	4.00e-02
8	170	6	1.30e-01	6	8.00e-02	4	7.00e-02
16	594	10	5.20e-01	6	3.60e-01	4	2.90e-01
32	2210	18	4.21e+00	5	2.48e+00	4	2.15e+00
64	8514	34	5.35e+01	4	2.36e+01	4	2.15e+01

$(M_\infty = 0.9)$

		P_0		P_1		P_2	
n	dim	#it	CPU in s	#it	CPU in s	#it	CPU in s
4	54	4	7.00e-02	4	3.00e-02	4	4.00e-02
8	170	6	1.30e-01	5	8.00e-02	5	7.00e-02
16	594	10	5.20e-01	5	3.20e-01	4	3.10e-01
32	2210	18	4.20e+00	5	2.37e+00	4	2.28e+00
64	8514	34	5.53e+01	5	2.29e+01	4	2.25e+01

TABLE 5.3

Performance of DCP for cost function parameters $\eta_1 = 1.0$, $\eta_2 = 0.0$.

$(M_\infty = 0.0)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	12	1.70e-01	4.4219e-01	4	5.00e-02	8.9447e-03
8	170	12	2.70e-01	4.8270e-01	4	8.00e-02	7.2777e-03
16	594	12	1.03e+00	4.8478e-01	4	3.10e-01	6.8858e-03
32	2210	11	6.98e+00	4.8348e-01	4	2.69e+00	6.7893e-03
64	8514	10	6.89e+01	4.8118e-01	3	2.76e+01	6.7653e-03

$(M_\infty = 0.1)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	12	1.60e-01	4.3988e-01	4	5.00e-02	9.2148e-03
8	170	12	2.70e-01	4.8033e-01	4	8.00e-02	7.5113e-03
16	594	12	1.02e+00	4.8238e-01	4	3.20e-01	7.1105e-03
32	2210	11	6.92e+00	4.8109e-01	4	2.69e+00	7.0118e-03
64	8514	10	6.99e+01	4.7879e-01	3	2.79e+01	6.9873e-03

$(M_\infty = 0.5)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	11	1.50e-01	3.7660e-01	4	5.00e-02	1.9846e-02
8	170	11	2.70e-01	4.1315e-01	4	8.00e-02	1.6904e-02
16	594	10	9.20e-01	4.1261e-01	4	3.10e-01	1.6202e-02
32	2210	10	6.58e+00	4.1312e-01	4	2.65e+00	1.6028e-02
64	8514	9	6.65e+01	4.1044e-01	4	3.06e+01	1.5985e-02

$(M_\infty = 0.9)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	10	1.40e-01	2.4531e-01	10	9.00e-02	2.4010e-01
8	170	9	2.20e-01	2.1838e-01	10	1.50e-01	2.2788e-01
16	594	9	8.50e-01	2.1129e-01	9	5.00e-01	2.2484e-01
32	2210	9	6.16e+00	2.0949e-01	9	3.91e+00	2.2409e-01
64	8514	9	6.53e+01	2.0904e-01	9	4.31e+01	2.2390e-01

TABLE 5.4

Performance of DCP for cost function parameters $\eta_1 = 1.0$, $\eta_2 = -1.0$.

$(M_\infty = 0.0)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	13	2.90e-01	3.8568e-01	12	1.60e-01	3.5942e-01
8	170	17	3.70e-01	4.8401e-01	17	2.20e-01	4.9946e-01
16	594	20	1.62e+00	5.2462e-01	20	9.40e-01	5.5028e-01
32	2210	21	1.17e+01	5.3696e-01	21	7.19e+00	5.6899e-01
64	8514	21	1.19e+02	5.4020e-01	21	7.51e+01	5.7651e-01

$(M_\infty = 0.1)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	13	1.80e-01	3.8366e-01	12	1.00e-01	3.5812e-01
8	170	17	3.70e-01	4.8193e-01	17	2.30e-01	4.9748e-01
16	594	20	1.58e+00	5.2247e-01	20	9.40e-01	5.4815e-01
32	2210	21	1.13e+01	5.3484e-01	21	6.94e+00	5.6681e-01
64	8514	21	1.13e+02	5.3811e-01	21	7.28e+01	5.7432e-01

$(M_\infty = 0.5)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	12	1.70e-01	3.3273e-01	11	1.00e-01	3.2583e-01
8	170	15	3.40e-01	4.2961e-01	15	2.00e-01	4.4778e-01
16	594	17	1.39e+00	4.6841e-01	17	8.50e-01	4.9393e-01
32	2210	18	9.90e+00	4.8126e-01	18	6.18e+00	5.1115e-01
64	8514	19	1.04e+02	4.8552e-01	18	6.56e+01	5.1815e-01

$(M_\infty = 0.9)$

		P_1			P_2		
n	dim	#it	CPU in s	conv.rate	#it	CPU in s	conv.rate
4	54	14	1.90e-01	3.6261e-01	14	1.10e-01	3.7223e-01
8	170	15	3.40e-01	4.1422e-01	16	2.20e-01	4.2473e-01
16	594	16	1.30e+00	4.3635e-01	16	7.90e-01	4.4671e-01
32	2210	16	9.07e+00	4.4537e-01	16	5.65e+00	4.5569e-01
64	8514	16	9.34e+01	4.4926e-01	16	6.00e+01	4.5958e-01

TABLE 5.5
LMA Prediction of asymptotic convergence rates for the DCP.

($\eta_1 = 1.0$ in all cases)

	P_1		P_2	
M_∞	$\eta_2 = 0.0$	$\eta_2 = -1.0$	$\eta_2 = 0.0$	$\eta_2 = -1.0$
0.0	0.5000	0.5781	0.0	0.5781
0.1	0.4975	0.5757	0.0	0.5757
0.5	0.4286	0.5124	0.0	0.5124
0.9	0.1597	0.2723	0.0	0.2723

REFERENCES

[1] E. ARIAN, *On the coupling of aerodynamic and structural design*, Journal of Computational Physics, 135 (1997), pp. 83–96.

[2] E. ARIAN AND S. TA'ASAN, *Analysis of the Hessian for Aerodynamic Optimization: Inviscid Flow*, ICASE Report No. 96–28, Institute for Computer Applications in Science and Engineering (ICASE), MS 403, NASA LaRC, Hampton, VA 23681-2199, 1996. To appear in Computers & Fluids.

[3] E. ARIAN AND V. VATSA, *A Preconditioning Method for Shape Optimization Governed by the Euler Equations*, ICASE Report No. 98–14, Institute for Computer Applications in Science and Engineering (ICASE), MS 403, NASA LaRC, Hampton, VA 23681-2199, 1996. To appear in International Journal of Computational Fluid Dynamics.

[4] A. BATTERMANN, *An Indefinite Preconditioner for Karush–Kuhn–Tucker Systems Arising in Optimal Control Problems*, tech. report, Universität Trier, Fachbereich IV, Abt. Mathematik, 54286 Trier, Germany, 1999. To appear.

[5] A. BATTERMANN AND M. HEINKENSCHLOSS, *Preconditioners for Karush–Kuhn–Tucker Systems Arising in the Optimal Control of Distributed Systems*, in Optimal Control of Partial Differential Equations, Vorau 1996, W. Desch, F. Kappel, and K. Kunisch, eds., Birkhäuser Verlag, Basel, Boston, Berlin, 1996, pp. 15–32.

[6] S. L. CAMPBELL, I. C. F. IPSEN, C. T. KELLEY, AND C. D. MEYER, *GMRES and the Minimal Polynomial*, BIT, 36 (1996), pp. 664–675.

[7] J. CONWAY, *A Course in Functional Analysis*, Springer–Verlag, New York, 1990.

[8] R. S. DEMBO, S. C. EISENSTAT, AND T. STEIHAUG, *Inexact Newton methods*, SIAM J. Numer. Anal., 19 (1982), pp. 400–408.

[9] R. W. FREUND, *Preconditioning of Symmetric, but Highly Indefinite Linear Systems*, Tech. Report 97–3–03, Bell Laboratories, 700 Mountain Avenue, Murray Hill, New Jersey 07974–0636, 1997.

[10] P. E. GILL, W. MURRAY, D. B. PONCELEÓN, AND M. A. SAUNDERS, *Preconditioners for indefinite systems arising in optimization*, SIAM J. Matrix Anal. Appl, 13 (1992), pp. 292–311.

[11] G. H. GOLUB AND C. F. VAN LOAN, *Matrix Computations*, The Johns Hopkins University Press, Baltimore and London, 1989.

[12] P. W. HEMKER, *Lecture Notes on Defect Correction*, Tech. Report Lecture Notes, Centrum voor Wiskunde en Informatica (CWI), P.O. Box 94079, NL – 1090 GB Amsterdam, 1997.

[13] C. HIRSCH, *Numerical Computation of Internal and External Flows, Vol. 2*, John Wiley & Sons, Inc., New York, 1988.

[14] C. T. KELLEY, *Iterative Methods for Linear and Nonlinear Equations*, SIAM, Philadelphia, 1995.

[15] C. T. KELLEY AND E. W. SACHS, *Quasi-Newton methods and unconstrained optimal control problems*, SIAM J. on Control and Optimization, 25 (1987), pp. 1503–1517.

[16] ——, *Multilevel algorithms for constrained compact fixed point problems*, SIAM J. Scientific and Stat. Computing, 15 (1994), pp. 645–667.

[17] A. KLAWONN, *An Optimal Preconditioner for a Class of Saddle Point Problems with a Penalty Term*, SIAM J. Sci. Comput., 19 (1998), pp. 540–552.

[18] ——, *Block–Triangular Preconditioners for Saddle Point Problems with a Penalty Term*, SIAM J. Sci. Comput., 19 (1998), pp. 5172–5184.

[19] M. LAUMEN AND E. W. SACHS, *Concepts of Newton and quasi–Newton methods for optimal shape design problems*, Control and Cybernetics, 25 (1996), pp. 895–913.

[20] D. G. LUENBERGER, *Optimization by vector space methods*, John Wiley & Sons, Inc., New York, 1969.

[21] M. RENARDY AND R. C. ROGERS, *An Introduction to Partial Differential Equations*, Springer Verlag, New York, 1993.

[22] T. RUSTEN AND R. WINTHER, *A preconditioned iterative method for saddlepoint problems*, SIAM J. Matrix Anal. Appl., 13 (1992), pp. 887–904.

[23] Y. SAAD AND M. H. SCHULTZ, *GMRES: A generalized minimal residual algorithm for solving nonsymmetric linear systems*, SIAM J. Sci. Stat. Comput., 7 (1986), pp. 856–869.

[24] H. J. STETTER, *The Defect Correction Principle and Discretization Methods*, J. Numer. Math., 29 (1978), pp. 425–443.

[25] J. STOER, *Solution of large linear systems of equations by conjugate gradient type methods*, in Mathematical Programming, The State of the Art, A. Bachem, M. Grötschel, and B. Korte, eds., Springer Verlag, Berlin, Heidelberg, New York, 1983.

[26] D. SYLVESTER AND A. WATHEN, *Fast iterative solution of stabilized Stokes systems part I: Using simple diagonal preconditioners*, SIAM J. Numer. Anal., 30 (1992), pp. 630–649.

[27] ——, *Fast iterative solution of stabilized Stokes systems part II: Using general block preconditioners*, SIAM J. Numer. Anal., 31 (1994), pp. 1352–1414.

[28] R. A. TAPIA, *Diagonalized multiplier methods and quasi-Newton methods for equality constrained optimization*, J. Optim. Theory Appl., 22 (1977), pp. 135–194.

[29] A. WATHEN, B. FISCHER, AND D. SYLVESTER, *The convergence rate of the minimal residual method for the Stokes problem*, Numer. Math., 71 (1995), pp. 121–134.

[30] E. ZEIDLER, *Nonlinear Functional Analysis and its Applications I*, Springer Verlag, New York, Berlin, Heidelberg, Tokyo, 1986.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. AGENCY USE ONLY *(Leave blank)*

2. REPORT DATE
February 1999

3. REPORT TYPE AND DATES COVERED
Contractor Report

4. TITLE AND SUBTITLE
Approximation of the Newton Step by a Defect Correction Process

5. FUNDING NUMBERS
C NAS1-97046
WU 505-90-52-01

6. AUTHOR(S)
E. Arian
A. Batterman
E.W. Sachs

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Institute for Computer Applications in Science and Engineering
Mail Stop 403, NASA Langley Research Center
Hampton, VA 23681-2199

8. PERFORMING ORGANIZATION REPORT NUMBER
ICASE Report No. 99-12

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
National Aeronautics and Space Administration
Langley Research Center
Hampton, VA 23681-2199

10. SPONSORING/MONITORING AGENCY REPORT NUMBER
NASA/CR-1999-209099
ICASE Report No. 99-12

11. SUPPLEMENTARY NOTES
Langley Technical Monitor: Dennis M. Bushnell
Final Report
To be submitted to the SIAM Journal of Optimization.

12a. DISTRIBUTION/AVAILABILITY STATEMENT
Unclassified Unlimited
Subject Category 64
Distribution: Nonstandard
Availability: NASA-CASI (301) 621-0390

12b. DISTRIBUTION CODE

13. ABSTRACT *(Maximum 200 words)*

In this paper, an optimal control problem governed by a partial differential equation is considered. The Newton step for this system can be computed by solving a coupled system of equations. To do this efficiently with an iterative defect correction process, a modifying operator is introduced into the system. This operator is motivated by local mode analysis. The operator can be used also for preconditioning in GMRES. We give a detailed convergence analysis for the defect correction process and show the derivation of the modifying operator. Numerical tests are done on the small disturbance shape optimization problem in two dimensions for the defect correction process and for GMRES.

14. SUBJECT TERMS
optimal control governed by PDEs, iterative methods, defect correction, GMRES, preconditioning, Newton step, SQP

15. NUMBER OF PAGES
35

16. PRICE CODE
A03

17. SECURITY CLASSIFICATION OF REPORT	18. SECURITY CLASSIFICATION OF THIS PAGE	19. SECURITY CLASSIFICATION OF ABSTRACT	20. LIMITATION OF ABSTRACT
Unclassified	Unclassified		